Nº 12-001-X au catalogue

Techniques d'enquête

Décembre 2008

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca. Vous pouvez également communiquer avec nous par courriel à infostats@statcan.gc.ca ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

Comment accéder à ce produit ou le commander

Le produit nº 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.gc.ca et de choisir la rubrique « Publications ».

Ce produit nº 12-001-X au catalogue est aussi disponible en version imprimée standard au prix de 30 $CAN l'exemplaire et de 58 $CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 $CAN	12 $CAN
Autres pays	10 $CAN	20 $CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
- Télécopieur (Canada et États-Unis) 1-877-287-4369
- Courriel infostats@statcan.gc.ca
- Poste
  Statistique Canada
  Finances
  Immeuble R.-H.-Coats, 6e étage
  150, promenade Tunney's Pasture
  Ottawa (Ontario) K1A 0T6
- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « À propos de nous » > « Offrir des services aux Canadiens ».

Statistique Canada
Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Décembre 2008

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2008

Tous droits réservés. Le contenu de la présente publication électronique peut être reproduit en tout ou en partie, et par quelque moyen que ce soit, sans autre permission de Statistique Canada, sous réserve que la reproduction soit effectuée uniquement à des fins d'étude privée, de recherche, de critique, de compte rendu ou en vue d'en préparer un résumé destiné aux journaux et/ou à des fins non commerciales. Statistique Canada doit être cité comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, il est interdit de reproduire le contenu de la présente publication, ou de l'emmagasiner dans un système d'extraction, ou de le transmettre sous quelque forme ou par quelque moyen que ce soit, reproduction électronique, mécanique, photographique, pour quelque fin que ce soit, sans l'autorisation écrite préalable des Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Décembre 2008

N° 12-001-XIF au catalogue
ISSN 1712-5685

N° 12-001-XPB au catalogue
ISSN 0714-0045

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-X)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Techniques d'enquête

Une revue éditée par Statistique Canada

Volume 34, numéro 2, décembre 2008

Table des matières

Dans ce numéro 141

Article Sollicité Waksberg

Mary E. Thompson
Enquêtes internationales : motifs et méthodologies 145

Articles Réguliers

Joanne Pascale et Alice McGee
Utilisation du codage du comportement pour évaluer l'efficacité de l'interview avec rétroinformation 159

Jing Xu, Jun Shao, Mari Palta et Lin Wang
Imputation pour les non-réponses non monotones dépendant de la dernière valeur dans les enquêtes longitudinales 169

Robert G. Clark et Raymond L. Chambers
Calage adaptatif pour la prédiction de totaux de population finie 181

Lionel Qualité et Yves Tillé
Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée 193

Inho Park
Masquage des UPE et estimation de la variance dans les enquêtes complexes 203

Roberto Benedetti, Giuseppe Espa et Giovanni Lafratta
Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes 217

Mohammad G. Mostafa Khan, Niraj Nand et Nesar Ahmad
Détermination des bornes optimales de strate au moyen de la programmation dynamique 227

José A. Díaz-García et Liliana Ulloa Cortez
Optimisation multi-objective pour une répartition optimale dans l'échantillonnage stratifié multivarié 237

Piero Demetrio Falorsi et Paolo Righi
Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines 247

Danny Pfeffermann, Bénédicte Terryn et Fernando A.S. Moura
Estimation pour petits domaines sous un modèle à effets aléatoires en deux parties avec application à l'estimation de la littératie dans les pays en voie de développement 261

Remerciements 277

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 141-142
Statistique Canada, N° 12-001-X au catalogue

Dans ce numéro

Ce numéro de *Techniques d'enquête* s'ouvre sur le huitième article de la série d'articles annuels sollicités dédiée à Joseph Waksberg. Le comité de rédaction tient à remercier les membres du comité de sélection, composé de Sharon Lohr, présidente, et de Bob Groves, Leyla Mohadjer et Wayne Fuller, d'avoir choisi Mary Thompson comme auteur de l'article du prix Waksberg de cette année.

Dans son article intitulé « Enquêtes internationales : motifs et méthodologies », Mary Thompson examine les défis que pose l'organisation et la collecte des données dans le cas d'enquêtes menées dans plusieurs pays simultanément, et se penche sur les problèmes que présentent les analyses qui comparent les différents pays dans lesquels les enquêtes sont réalisées. Elle donne plusieurs exemples d'enquêtes internationales et illustre les défis qui se posent en examinant en plus grand détail l'enquête menée dans le cadre du projet International Tobacco Control. Elle examine également plusieurs méthodes différentes d'étalonnage des mesures et employées pour les comparaisons transculturelles.

Pascale et McGee présentent une étude sur l'utilisation de l'interview avec rétroinformation, technique utilisée dans de nombreuses enquêtes longitudinales pour « reporter » les données d'une vague à la suivante. À partir d'enregistrements d'interviews menées sur le terrain, les auteures utilisent le codage comportemental pour évaluer l'efficacité de l'interview avec rétroinformation. L'article apporte des éclaircissements intéressants sur l'influence du type de données reportées sur la façon dont les intervieweurs posent les questions.

L'article de Xu, Shao, Palta et Wang porte sur l'imputation des valeurs manquantes dans les enquêtes longitudinales lorsque le profil de non-réponse n'est pas monotone. Les auteurs supposent que le mécanisme de non-réponse de la période en cours dépend seulement de la dernière valeur de la variable à imputer et que cette variable suit une chaîne de Markov. L'imputation est effectuée au moyen d'une série de modèles de régression non paramétriques. Une méthode bootstrap est utilisée aux fins de l'estimation de la variance. La méthode est illustrée au moyen de données simulées ainsi que de données réelles.

Dans le contexte du cadre de prédiction, Clark et Chambers proposent une approche de calage adaptive pour la sélection d'un ensemble approprié de renseignements auxiliaires. Ils appliquent leur méthode à une vaste gamme de modèles. Les résultats d'une étude en simulation sont présentés pour confirmer le bon rendement des méthodes proposées.

L'estimation de la variance d'estimateurs du changement entre deux périodes successives d'une enquête répétée est étudiée par Qualité et Tillé. Dans cet article, on tient compte, entre autres, du plan de sondage qui utilise un panel, de la non-réponse totale, de la réduction des poids de sondage due aux valeurs aberrantes et du calage. La méthodologie proposée est appliquée à l'Enquête suisse sur la valeur ajoutée.

Dans son article, Park se penche sur les difficultés que pose aux analystes l'utilisation de données à grande diffusion qui ont été modifiées pour assurer la protection des renseignements personnels. Il examine plus particulièrement l'effet de la permutation d'unités primaires d'échantillonnage (UPE), à laquelle on procède couramment pour protéger l'identité des répondants aux enquêtes, sur le calcul des variances. Il propose un nouvel algorithme de permutation des UPE et compare ses effets sur l'estimation de la variance, théoriquement et empiriquement, au moyen de plusieurs méthodes existantes.

Benedetti, Espa et Lafratta proposent un processus séquentiel de stratification d'une population finie. Ce processus sert à obtenir une stratification multivariée selon une approche fondée sur l'élaboration d'un arbre. Au moyen de ce processus, ils produisent des partitions de plus en fines de la population jusqu'à ce que la différence entre les tailles d'échantillon optimales obtenue par les deux étapes consécutives soit inférieure à un niveau prédéterminé. L'approche proposée est appliquée à l'Enquête sur la structure des exploitations agricoles en Italie.

Khan, Nand et Ahmed se penchent sur le problème de la détermination des bornes optimales de strate comme problème de détermination des largeurs optimales de strate. Ils le formulent comme un problème de programmation mathématique et, pour le résoudre, étendent l'approche de programmation dynamique de Bühler et Deutler (1975). L'article est une extension de cette approche de programmation dynamique à des variables d'intérêt suivant une distribution triangulaire et une distribution normale standard. Une petite étude par simulation compare la méthode proposée à la méthode de la racine carrée cumulative (f) de Dalenius et Hodges (1959) révélant des gains d'efficacité.

Dans leur article, Díaz-García et Cortez examinent le problème de la répartition dans l'échantillonnage stratifié multivarié. Ils expriment ce problème comme un problème non linéaire d'optimisation matricielle d'entiers relatifs sous la contrainte d'une fonction de coût ou d'une taille d'échantillon donnée. Ils appliquent leur méthode en utilisant des données tirées d'une enquête sur les forêts.

Enfin, la conférence satellite sur l'estimation pour petits domaines de 2007 (EPD 2007) de l'IIS (Institut international de statistique) a produit une myriade d'articles portant sur les dernières techniques d'estimation pour petits domaines. Plusieurs de ces articles ont été soumis à *Techniques d'enquête*. Deux d'entre eux sont les derniers articles du présent numéro; nous nous attendons à ce que d'autres soient publiés dans les numéros à venir.

Falorsi et Righi ont élaboré une stratégie d'échantillonnage pour obtenir les tailles d'échantillon prévues pour les domaines sujets à des erreurs d'échantillonnage prédéterminées, particulièrement lorsque la classification croisée des variables définissant les différentes partitions donnerait un nombre de strates supérieur à la taille globale de l'échantillon. La méthode proposée présente l'avantage de la faisabilité des calculs et de la mise en œuvre d'une stratégie d'estimation pour petits domaines qui comprend à la fois le plan d'échantillonnage et l'estimation et améliore l'efficacité des estimateurs directs par domaine.

Dans le dernier article de ce numéro, Pfeffermann, Terryn et Moura se penchent sur des situations où la valeur de la réponse cible est soit zéro, soit une observation d'une loi continue, par exemple lors de l'évaluation de compétences en littératie dont le résultat possible est soit zéro, indiquant que la personne est analphabète, soit un score positif mesurant le niveau de littératie. Les méthodes disponibles, toutefois, ne conviennent pas à ces types de données étant donné la distribution mixte des réponses, un pic important à zéro étant juxtaposé à une distribution continue pour le reste des réponses. Les auteurs élaborent un modèle à effets aléatoires en deux parties approprié et ils montrent comment estimer le modèle pour en évaluer la qualité de l'ajustement et comment calculer les estimateurs pour petits domaines d'intérêt et en mesurer la précision.

Et finalement, nous sommes heureux d'informer les lecteurs et auteurs que *Techniques d'enquête* est maintenant citée dans le ISI Web of knowledge, ce qui inclue le Current Contents/Social and Behaviorial Sciences, Social Sciences Citation Index, and the Science Citation Index Expanded depuis le numéro de juin 2007.

Harold Mantel, Rédacteur en chef délégué

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 143-144
Statistique Canada, N° 12-001-X au catalogue

Série Waksberg d'articles sollicités

La revue *Techniques d'enquête* a mis sur pied une série de communications sollicitées en l'honneur de Joseph Waksberg, qui a fait de nombreuses contributions importantes à la méthodologie d'enquête. Chaque année, un éminent chercheur est choisi pour rédiger un article pour la série de communications sollicitées de Waksberg. L'article examine les progrès et l'état actuel d'un thème important dans le domaine de la méthodologie d'enquête et reflète l'agencement de théorie et de pratique caractéristique des travaux de Waksberg. L'auteur reçoit une prime en argent qui provient d'une bourse de Westat, en reconnaissance des contributions de Joe Waksberg pendant ses nombreuses années de collaboration avec Westat. L'administration financière de la bourse est assurée par l'American Statistical Association. La liste des gagnants précédents est présentée ci-bas. Leurs articles sont déjà parus dans la revue *Techniques d'enquête*.

Gagnants précédents du prix Waksberg :

Gad Nathan (2001)
Wayne A. Fuller (2002)
Tim Holt (2003)
Norman Bradburn (2004)
J.N.K. Rao (2005)
Alastair Scott (2006)
Carl-Erik Särndal (2007)
Mary Thompson (2008)

Nominations :

L'auteur de l'article Waksberg de 2010 sera sélectionné par un comité de quatre personnes désignées par *Techniques d'enquête* et l'American Statistical Association. Les candidatures ou les suggestions de sujets doivent être envoyées à Leyla Mojadjer, présidente du comité, par courriel à MOHADJL1@WESTAT.com. Les candidatures et les suggestions de sujets doivent être reçues d'ici au 27 février 2009.

Article sollicité Waksberg 2008

Auteur : Mary Thompson

Mary Thompson est professeure de statistique à l'Université de Waterloo à Waterloo en Ontario. Elle est co-directrice du Survey Research Centre de l'Université de Waterloo et directrice du Centre de gestion des données pour le Projet international d'évaluation de la lutte antitabac. La théorie de l'estimation, particulièrement pour les enquêtes complexes a été au centre de ses intérêts de recherche au cours des années. Elle est l'auteure d'une monographie intitulée Theory of Sample Surveys, publiée en 1997. Récemment, elle s'est intéressée à l'analyse des données d'enquêtes avec des modèles à variables latentes dans le domaine des sciences sociales et des sciences de la santé.

Membres du comité de sélection de l'article Waskberg (2008-2009)

Leyla Mojadjer (Présidente), *Westat*
Wayne A. Fuller, *Iowa State University*
Daniel Kasprzyk, *Mathematica Policy Research*
Betsy Martin

Présidents précédents :

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Bob Groves (2007-2008)

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 145-157
Statistique Canada, N° 12-001-X au catalogue

Enquêtes internationales : motifs et méthodologies

Mary E. Thompson [1]

Résumé

Le contexte de la discussion est la fréquence croissante des enquêtes internationales, dont fait partie l'International Tobacco Control (ITC) Policy Evaluation Project, qui a débuté en 2002. Les enquêtes ITC nationales, qui sont longitudinales, ont pour but d'évaluer les effets des mesures stratégiques adoptées dans divers pays en vertu de la Convention-cadre pour la lutte antitabac de l'OMS. Nous examinons et illustrons les défis que posent l'organisation, la collecte des données et l'analyse des enquêtes internationales. L'analyse est une raison qui motive de plus en plus la réalisation d'enquêtes transculturelles à grande échelle. La difficulté fondamentale de l'analyse est de discerner la réponse réelle (ou le manque de réponse) aux changements de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Deux problèmes ayant trait à l'analyse statistique sont examinés. Le premier est celui de savoir quand et comment analyser des données regroupées provenant de plusieurs pays, afin de renforcer des conclusions qui pourraient être généralement valides. Bien que cela paraisse simple, dans certains cas les avis sont partagés quant à la mesure dans laquelle ce regroupement est possible et raisonnable. Selon certains, les modèles à effets aléatoires sont conceptuellement utiles pour les comparaisons formelles. Le deuxième problème consiste à trouver des modèles de mesure applicables à diverses cultures et à divers modes de collecte de données qui permettent l'étalonnage des réponses continues, binaires et ordinales, ainsi que la production de comparaisons dont ont été éliminés les effets extérieurs. Nous constatons que les modèles hiérarchiques offrent un moyen naturel de relâcher les exigences d'invariance du modèle entre les groupes.

Mots clés : Enquêtes internationales ; enquêtes longitudinales ; analyse de données d'enquête ; effets aléatoires ; effets du mode de collecte des données ; modèles hiérarchiques ; modèles de mesure.

1. Introduction

J'ai choisi comme thème de mon exposé les enquêtes internationales, parce que les quelques dernières années, une grande part de mon activité a été consacrée à l'une de ces enquêtes, l'International Tobacco Control survey, et parce qu'il existe certains recoupements intéressants avec les domaines auxquels s'est intéressé Joseph Waksberg, en particulier les bases de sondage pour les enquêtes téléphoniques et les effets de la stratification avec des taux d'échantillonnage très variables.

L'article débute par une discussion des raisons qui motivent les enquêtes internationales et par certains exemples. Puis, il traite des défis que posent l'organisation, la collecte des données et l'analyse. Enfin, il aborde deux problèmes qu'il convient de résoudre dans le cadre de l'analyse, à savoir i) celui de la théorie de l'échantillonnage et du regroupement de données provenant de plusieurs pays et ii) celui de la mesure dans le contexte de différents modes de collecte des données et de différentes cultures.

La première grande enquête internationale a été l'Enquête mondiale sur la fécondité (EMF), menée durant les années 1970 par l'entremise de l'Institut international de statistique et financée par l'Agency for International Development des États-Unis et d'autres organismes parrains. Il s'agissait d'une enquête ponctuelle très ambitieuse. En fin de compte, l'EMF a été réalisée auprès de 330 000 femmes dans 61 pays, au coût d'environ 50 millions de dollars. Elle a fourni aux pays participants d'importantes données comparatives sur les tailles des familles et a abouti à l'adoption d'une politique de planification démographique dans plusieurs pays participants. Elle a également donné naissance à des centaines de projets analytiques, dont certaines études méthodologiques inédites, et a jeté les fondations de la méthodologie d'enquête internationale, particulièrement dans les pays en voie de développement (Verma, Scott et O'Muircheartaigh 1980 ; Cleland et Verma 1989).

Un autre exemple bien connu est le Programme international pour le suivi des acquis des élèves, un projet de l'Organisation de coopération et de développement économiques, qui a été lancé en 2000. Le PISA est une enquête continue, exécutée tous les trois ans, auprès des jeunes de 15 ans dans les pays développés. Sa portée s'accroît, 67 pays devant, en principe, y participer en 2009. Les résultats permettent aux pays de surveiller le succès de leurs programmes d'enseignement en ce qui concerne l'acquisition de compétences en communication verbale et en compréhension de texte à contenu quantitatif.

La Global Youth Tobacco Survey (GYTS) est une enquête ponctuelle parrainée par l'Organisation mondiale de la santé et les Centers for Disease Control and Prevention qui a débuté en 2002. Elle a pour cible les jeunes de 13 à 15 ans des pays en voie de développement et, en 2004, des données avaient été recueillies dans 129 pays. L'objectif est de mesurer l'adoption du tabagisme chez les jeunes et la sensibilisation aux risques connexes pour la santé.

1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. Courriel : methomps@uwaterloo.ca.

L'Enquête sociale européenne (ESE 2008) est une « enquête sociale appuyée par les universités » réalisée dans plus de 30 pays, financée par des organismes européens et nationaux, et conçue pour « expliquer l'interaction entre les institutions européennes en pleine évolution et les attitudes, les croyances et les comportements des diverses populations ».

Si le recours à des enquêtes locales et nationales est en hausse partout, il en est de même de la fréquence des enquêtes internationales, exécutées par des organismes internationaux, des organisations non gouvernementales et des entreprises du secteur privé. Cet essor semble s'inscrire dans une tendance vers une gouvernance mondiale et une préoccupation pour la santé et le bien-être des populations.

J'ai vu les objectifs des enquêtes internationales classés sous les rubriques de l'épidémiologie, de la surveillance et de l'évaluation des effets des politiques. De toute évidence, ces classifications se chevauchent. On peut soutenir que le PISA, la GYTS et l'ESE rentrent dans les catégories de la surveillance, parce que leurs données ne sont reliées qu'indirectement aux interventions. Par contre, l'EMF comporte un aspect d'évaluation directe dans les pays qui ont mis sur pied des programmes de planification familiale. L'International Tobacco Control (ITC) survey, dont il sera question plus loin dans cette section, est l'une des rares enquêtes dont l'objectif principal est l'évaluation.

Outre les préoccupations d'ordre scientifique, l'un des rôles importants d'une enquête internationale consiste à éveiller l'intérêt des gouvernements ; elle leur offre un moyen de participer à l'élaboration de politiques mondiales, même quand ils font face à des obstacles politiques et économiques.

Aux chercheurs, les enquêtes internationales permettent de comparer les populations de divers pays, d'interpréter les différences, voire même de mieux en comprendre les causes et les effets – habituellement avec l'objectif sous-jacent d'améliorer les conditions existantes et d'appuyer l'élaboration des politiques.

L'International Tobacco Control Policy Evaluation Project (projet ITC) a été lancé par Geoffrey T. Fong, Ph.D., du Département de psychologie à l'Université de Waterloo, et par des collaborateurs partout dans le monde (Fong, Cummings, Borland, Hastings, Hyland, Giovino, Hammond et Thompson 2006 ; Thompson, Fong, Hammond, Boudreau, Dreizen, Hyland, Borland, Cummings, Hastings, Siahpush, Mackintosh et Laux 2006). L'élan a été donné par la Convention-cadre sur la lutte antitabac (CCLA) de l'OMS, qui a été adoptée en mai 2003 et ratifiée par plus de 150 pays. En ratifiant le traité, les pays participants ont promis des mesures stratégiques de lutte antitabac, telles que l'apposition d'étiquettes portant des mises en garde musclées concernant le danger pour la santé, l'interdiction de la publicité au profit des cigarettes et l'interdiction de fumer dans les lieux publics. Parce que ces mesures requièrent une législation nationale, le moment où elles sont adoptées et la façon dont elles le sont varient. Par exemple, en décembre 2000, le Canada a commencé à utiliser des étiquettes de mise en garde explicites, créant un précédent international en ce qui concerne la taille de l'étiquette (plus de 50 % de la surface de l'emballage) et l'utilisation d'images aux couleurs vives. Depuis, quelques autres pays ont opté pour la même pratique, tandis que d'autres ont adopté des lois prévoyant l'utilisation d'avertissements textuels bien en vue. Pour des renseignements sur la situation actuelle de la réglementation concernant les mises en garde relatives à la santé dans le monde, consulter ITC (2008). Le rapport MPower (OMS 2008) décrit l'environnement stratégique mondial concernant la lutte antitabac et six stratégies recommandées pour la CCLA.

Le but du projet ITC est d'essayer de trouver des stratégies qui réduisent effectivement le taux d'adoption du tabagisme et qui aident les jeunes qui fument déjà à arrêter de le faire. S'ajoute à cela l'objectif ambitieux d'essayer d'expliquer comment fonctionnent vraiment les stratégies efficaces. L'équipe de recherche comprend des psychologues sociaux et des spécialistes du marketing social, ainsi que des épidémiologistes et des économistes.

En septembre 2008, 17 pays réalisaient des enquêtes dans le cadre du projet ITC et d'autres s'ajouteront vraisemblablement à la liste. Les enquêtes ont débuté en 2002 au Canada, aux États-Unis, au Royaume-Uni et en Australie. Cette année-là, dans chacun de ces quatre pays, environ 2 000 fumeurs adultes ont été sélectionnés par téléphone en utilisant une base de sondage par composition aléatoire (CA) stratifiée géographiquement, dont le fondement scientifique a pour origine la fameuse méthode de Mitofsky-Waksberg (Waksberg 1978). Les fumeurs sélectionnés ont été interviewés une semaine ou deux plus tard, puis ont fait l'objet d'un suivi chaque année depuis, qu'ils aient continué ou non de fumer. La sixième vague de l'enquête ITC-Quatre pays s'est achevée en février 2008.

Comme une taille d'échantillon suffisante est nécessaire pour évaluer les effets des mesures introduites entre les vagues de l'enquête, les participants qui décrochent à chaque vague sont remplacés par une cohorte de nouvelles personnes. Dans le cadre de l'enquête ICT-Quatre pays, à chaque vague, les nouveaux participants ont été sélectionnées en utilisant le même plan de sondage qu'à la première vague, sans essayer de rechercher des personnes ayant les mêmes caractéristiques que celles ayant décroché. À chaque vague, la construction des poids est, en fait, exécutée séparément pour chaque cohorte, en procédant à un rajustement pour tenir compte des différences d'attrition selon la région et le groupe âge-sexe. Ce plan de sondage nous a permis de discerner des effets de « durée de la présence dans

l'échantillon », si bien que cette caractéristique est introduite en tant que variable explicative dans les modèles analytiques (Thompson, Boudreau et Driezen 2005).

Après 2002, les premières mesures stratégiques nationales qui ont été prises étaient l'interdiction de la publicité et une amélioration des étiquettes de mise en garde entre les première et deuxième vagues au Royaume-Uni, ainsi que l'adoption d'étiquettes de mise en garde explicites entre les quatrième et cinquième vagues en Australie. L'enquête ITC-Quatre pays constitue ce que l'on a parfois appelé une expérience naturelle ou quasi-expérience (Cook et Campbell 1979), les pays où une politique particulière n'existe pas servant de contrôles externes ; en outre, l'aspect longitudinal du plan de sondage offre un contrôle interne. Le plan de sondage a été répété un certain nombre de fois, pour d'autres groupes de pays.

Par exemple, très vite, il est devenu évident que l'Irlande serait le premier pays à adopter une loi antitabac nationale, laquelle est entrée en vigueur en mars 2004. Les collaborateurs du projet ITC ont réussi à mettre sur pied des enquêtes parallèles en Irlande et au Royaume-Uni avant que la loi n'entre en vigueur et à rendre visite aux mêmes personnes une année plus tard. De nouveau, les échantillons ont été sélectionnés nationalement à l'aide d'une base de sondage par composition aléatoire (CA). En tout, 755 fumeurs ont été interviewés lors des deux vagues en Irlande, et 411 au Royaume-Uni. Une constatation intéressante concerne l'appui pour une interdiction de l'usage du tabac dans les pubs (Fong, Hyland, Borland, Hammond, Hastings, McNeill, Anderson, Cummings, Allwright, Mulcahy, Howell, Clancy, Thompson, Connolly et Driezen 2006). La figure 1 donne les proportions de personnes appuyant ou appuyant fortement l'interdiction dans les bars et les pubs, dans les deux pays, selon la vague.

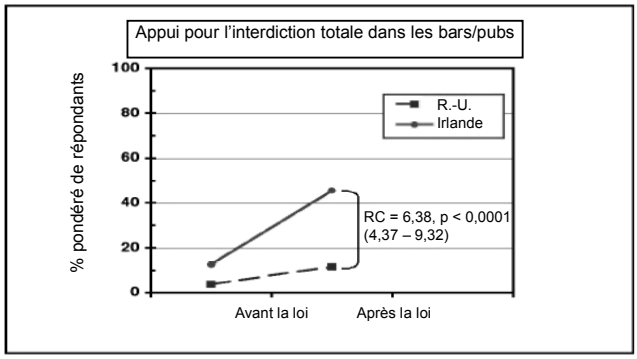


Figure 1 Appui pour la loi antitabac durant deux vagues de l'enquête ITC en Irlande et au Royaume-Uni

Dans l'échantillon ITC de fumeurs, l'accroissement de l'appui pour l'interdiction entre les deux vagues a été faible au Royaume-Uni et important en Irlande. En outre, l'enquête n'a donné aucune preuve que la réduction de l'usage du tabac dans les lieux publics était associée à un accroissement dans les lieux privés. Révélant l'acceptation générale de la loi antitabac par les fumeurs, les résultats de l'enquête ITC et d'autres résultats comparables ont facilité l'adoption de lois semblables en Écosse, en France, en Allemagne, dans le reste du Royaume-Uni et aux Pays-Bas. Une enquête ITC a été réalisée avant et après la mise en application de l'interdiction, en avril 2006, en Écosse, en utilisant le reste du Royaume-Uni comme contrôle, et les résultats ont été reproduits, sauf qu'à ce moment-là, l'appui dans le reste du Royaume-Uni avait augmenté considérablement (Hyland, Hassan, Higbee, Fong, Borland, Cummings, Thompson, Boudreau et Hastings 2008).

Le modèle utilisé pour le test était simple, à savoir un modèle EEG, où Y est une mesure binaire de l'appui pour l'interdiction, w désigne le pays, t représente le temps, le terme wt représente une interaction, et x est un vecteur de covariables fixes au niveau individuel :

$$\text{logit}[P(Y_t = 1 \mid w, x) = \alpha_0 + \alpha_1 w + \gamma t + \delta wt + x\beta,$$

$$\text{Corr}(Y_1, Y_2) = \rho.$$

Le coefficient δ représente la différence d'accroissement de l'appui dans les deux pays, et nous avons testé l'hypothèse H_0: $\delta = 0$. D'autres paramétrisations sont possibles, mais celle-ci a l'avantage de concorder avec le graphique de la figure 1, qui représente des proportions marginales ; la méthodologie est généralement reconnue et prise en charge par les logiciels pour données d'enquêtes complexes.

2. Défis

La réalisation d'une enquête internationale pose de nombreux défis. Les articles sur l'EMF publiés par Verma et coll. (1980) et par Cleland et Verma (1989) contiennent des discussions réfléchies à ce sujet qui ne sont guère dépassées aujourd'hui. À la présente section, en guise d'illustration, nous décrivons certains problèmes survenus durant l'organisation et la collecte des données de l'enquête ITC.

Contrairement à l'EMF, l'enquête ITC a été financée, en premier lieu, par des programmes nationaux de subventions, principalement les National Institutes of Health des États-Unis et les Instituts de recherche en santé du Canada. L'infrastructure centrale, ayant à sa tête M. Fong à l'Université de Waterloo et M. K. Michael Cummings au Roswell Park Cancer Institute à Buffalo, collabore directement avec les organismes des divers pays. Nous avons dû apprendre à travailler avec des groupes provenant de sociétés, de régimes politiques et de cultures très différents. Rien que les coûts et les budgets de l'enquête diffèrent

étonnamment d'un pays à l'autre. Quand les gouvernements participent au financement, ils ont leurs propres exigences et des ententes concernant la propriété des données doivent être négociées. Puisque les niveaux d'infrastructure et d'expertise peuvent différer assez bien d'un pays à l'autre, la coordination étroite qui existe dans le cas de l'enquête ITC-Quatre pays est difficile à reproduire à une plus grande échelle.

Par exemple, durant la première moitié de 2008, le travail sur le terrain a été effectué pour la troisième vague d'une enquête parallèle (l'enquête ITC-Asie du Sud-Est) en Thaïlande et en Malaisie, pays qui sont géographiquement proches et semblables à certains égards, mais qui diffèrent en ce qui a trait à de nombreux aspects. Du point de vue ethnique, la Thaïlande est assez homogène, alors que la Malaisie compte trois grands groupes ethniques et de nombreux groupes mineurs. En Thaïlande, plus de la moitié de la population vit dans les régions rurales, mais en Malaisie, la plupart de la population est urbaine et la mobilité résidentielle est forte. La Thaïlande a accumulé une grande expérience des enquêtes, y compris les études de cohorte, mais quand la première vague de l'enquête a démarré en 2005, la Malaisie entreprenait ce genre d'étude de cohorte pour la première fois. Nous avons essayé de recommander des plans d'échantillonnage parallèles dans les deux pays, mais nous avons dû faire des compromis. Par exemple, il a été découvert, au moment de la première vague, que les bases de sondage officielles étaient constituées, au niveau le plus bas, de blocs élémentaires de tailles différentes correspondant à des grappes de ménages. Cette différence a rendu l'échantillon de ménages plus dispersé en Malaisie, où les blocs étaient plus petits. La dispersion plus forte signifiait plus de travail et des coûts plus élevés. (Les effets de plan sont encore plus importants pour la Malaisie que pour la Thaïlande, à cause de la plus forte hétérogénéité au niveau des unités de premier degré.)

Un aspect important du projet consiste à essayer d'accroître la capacité d'exécution d'enquêtes sur la santé longitudinales dans des pays où ce genre de travail est relativement nouveau. Nous fournissons des protocoles détaillés, des manuels de formation et des modèles de saisie des données. Nous avons appris à insister davantage sur la détermination des compétences locales, particulièrement en statistique. La communication quotidienne se fait par courriels et par téléconférences. L'épuration des données finales et la construction des poids d'enquête se font normalement à l'Université de Waterloo, mais les équipes de certains pays désirent vivement participer à ces étapes des opérations.

Nous procédons à des enquêtes téléphoniques, avec sélection des participants par CA modifiée dans les quatre pays originaux, ainsi qu'en Irlande, en Corée du Sud, en France et en Allemagne et avec sélection à partir de l'échantillon de la National Health Survey en Nouvelle-Zélande, ainsi qu'à des enquêtes par interview sur place en Thaïlande, en Malaisie pour la première vague, en Chine, au Bangladesh, au Mexique et en Uruguay.

En Malaisie, pour la deuxième vague, nous avions l'intention de procéder à des interviews sur place, mais comme la reprise de contact ainsi que la sélection de nouveaux participants se sont avérées difficiles à cause d'une combinaison de facteurs, nous sommes passés à l'interview téléphonique dans la mesure du possible dans certaines régions. La possibilité de comparer les modes de collecte était limitée, mais dans le grand État principalement urbain de Selangor, 137 fumeurs interrogés à la première vague (qui n'avaient pas arrêté) ont été réinterviewés sur place, et 63 ont été réinterviewés par téléphone, ce qui rend possible certaines inférences provisoires. Pour la troisième vague, nous avons essayé de procéder à des interviews par téléphone et sur place dans certains des mêmes districts de recensement, en vue de permettre une meilleure évaluation des effets du mode de collecte des données, et cette étude est en cours. Parallèlement, la proportion de l'échantillon de fumeurs de l'enquête ITC-Malaisie interviewés sur place a diminué régulièrement, pour passer de 100 % à la première vague à 63,5 % à la deuxième, et à 44,4 % à la troisième. Pour la quatrième vague, nous prévoyons utiliser seulement la collecte par téléphone dans les États continentaux.

Pour la première vague de l'enquête, les Pays-Bas ont utilisé parallèlement un panel Internet et l'interview téléphonique avec CA, avec des tailles d'échantillon d'environ 400 et 1 800, respectivement. Cet exercice offrira la meilleure occasion jusqu'à présent de pouvoir tenir compte des effets de mode dans la modélisation. Ces effets ont fait l'objet de nombreux travaux de recherche récemment. Par exemple, certaines études ont montré que les personnes interviewées par téléphone choisissent les options extrêmes d'une échelle de Likert plus fréquemment que celles qui répondent par Internet (Wichers et Zenderink 2006 ; Bronner et Kuijlen 2007).

Aux Pays-Bas, l'échantillon de l'enquête en ligne est constitué de fumeurs sélectionnés aléatoirement à partir d'un grand panel polyvalent prérecruté d'environ 200 000 personnes créé par la firme TNS NIPO. L'échantillon de l'enquête téléphonique, sélectionné parmi les fumeurs abonnés à un service téléphonique par ligne terrestre, pourrait fort bien représenter une population de fumeurs différente. Le faible taux de réponse téléphonique montre clairement qu'aux Pays-Bas, les membres du public ne sont pas aussi réceptifs aux enquêtes téléphoniques que dans la plupart des autres pays participant au projet ITC. Nous avons souhaité que l'on demande à chaque groupe s'il pouvait être rejoint par l'autre mode, afin de pouvoir utiliser

des méthodes à base de sondage double (Lohr et Rao 2000) pour calculer les poids d'enquête appropriés. Nous modéliserons également la propension (Rosenbaum et Rubin 1984) à répondre par téléphone (disons), sachant les variables démographiques et les variables d'accessibilité, et nous tiendrons compte du score de propension dans les comparaisons des profils de réponse selon le mode.

Les taux de réponse varient beaucoup, même dans le cas de l'enquête ITC-Quatre pays, les taux de réponse et les taux de rétention étant les plus élevés en Australie et les plus faibles aux États-Unis. Cela compromet certainement la capacité de faire des comparaisons entre pays, en ce sens que nous ne pouvons comparer que les populations représentées par les répondants, c'est-à-dire ceux qui, dans chaque pays, répondraient s'ils étaient sollicités aux termes de notre protocole. La situation paraît un peu meilleure si nous décomposons les taux de réponse en leurs composantes. Par exemple, nous avons constaté, d'après les résultats des tentatives d'appel et ce que nous savons de l'utilisation croissante d'appareils de filtrage des appels, qu'il est beaucoup plus difficile de prendre et de reprendre contact avec les adultes américains qu'avec ceux résidant dans les trois autres pays. Cependant, une fois que le contact est établi, le taux d'accord ou de non-refus aux États-Unis (plus de 80 %) est très semblable à ceux observés dans les trois autres pays.

Nous nous heurtons à des problèmes de mesure, même pour des faits ordinaires, tels que les habitudes d'achat et de consommation de tabac. Dans certains pays, comme l'Inde, le Bangladesh et le Soudan, dont l'inclusion dans le projet est en cours de discussion, de nombreuses formes de tabac sont utilisées couramment. Dans les pays développés, rien que tenir à jour la liste des marques de cigarettes est un travail à temps plein. Vient aggraver la difficulté le fait que, quand nous demandons aux participants à l'enquête quelles sont leurs habitudes d'achat ou s'ils ont remarqué les publicités, nous leur demandons de se souvenir de ce qu'ils ont fait au cours des deux semaines précédentes ou au cours d'une période plus longue. Dans la majorité des cas, nous nous fions à des données autodéclarées, mais pour un certain nombre de raisons, l'autodéclaration pourrait ne pas être exacte.

En ce qui concerne les attitudes et les croyances, nous savons depuis toujours que les questions doivent être adaptées à la langue et au niveau de littératie des participants, mais nous avons néanmoins été surpris et brusquement rappelés à la réalité de constater une forte prévalence de non-réponses partielles dans les régions éloignées d'un des pays, ce qui donne à penser que les questions sur les attitudes et les croyances présentent de grandes difficultés. Dans le cas de l'enquête pilote réalisée en Inde, l'interview a duré, en moyenne, 1,5 heure par participant, bien que le questionnaire ait été raccourci et simplifié.

Les mesures psychosociales doivent être validées dans chaque culture et chaque langue. Ainsi, nous avons commencé à inclure une très courte échelle de dépression. Voici la version utilisée pour l'enquête ITC-Quatre pays.

- Au cours du dernier mois, avez-vous souvent éprouvé peu d'intérêt ou de plaisir à faire les choses ?
- Au cours du dernier mois, avez-vous souvent éprouvé le sentiment d'être abattu(e), déprimé(e) ou désespéré(e) ?
- Au cours de la dernière année, un médecin ou un autre fournisseur de soins de santé vous a-t-il dit que vous faisiez une dépression ?

Et voici la version que nous avons finalement adoptée pour la deuxième vague de l'enquête ITC-Chine, en suivant les conseils d'autres chercheurs qui ont signalé qu'ils avaient pu en valider une version.

Voici une liste d'états que vous pourriez avoir ressentis ou de comportements que vous pourriez avoir eus. Dites-moi à quelle fréquence vous vous êtes senti(e) de cette façon durant la dernière semaine.

1. Je n'avais pas envie de manger ; mon appétit n'était pas bon.
2. Je me sentais sans espoir au sujet de l'avenir.
3. Je me sentais triste.
4. J'avais l'impression que les gens ne m'aiment pas.

Ryder, Yang, Zhu, Yao, Yi, Heine et Bagby (2008) ont compilé les résultats d'une étude comparative fort intéressante de l'expression de la dépression.

La liste des problèmes de mesure continue. Même si la question est censée être la même dans les deux langues, il peut être difficile de trouver des équivalences. Nous nous efforçons d'obtenir une traduction de bonne qualité en utilisant un comité de la traduction ou en comparant des traductions indépendantes, mais nous devons souvent accepter un résultat imparfait. Par exemple, les traductions littérales de l'anglais au français ou à l'allemand sont généralement plus longues dans la langue d'arrivée et l'obtention d'une traduction facile à utiliser par téléphone requiert beaucoup de compétence. Thrasher, Quah, Borland, Awang, Sirirassamee, Boado, Miller, Watts et Dorantes (2008) décrivent une étude d'évaluation cognitive de certaines des questions les plus importantes qui a été réalisée dans plusieurs pays.

Il existe aussi des différences culturelles plus subtiles, particulièrement la mesure dans laquelle les répondants donnent une réponse socialement désirable. Nous avons relevé ce qui pourrait être une tendance plus prononcée à le

faire chez les Mexicains et chez les Canadiens anglophones. Johnson et Van de Vijver (2003), entre autres, ont examiné la possibilité que les différences entre pays, pour ce qui est des réponses socialement désirables, soient reliées aux « systèmes de valeurs culturelles, comme dans la dimension de l'individualisme/collectivisme » de Hofstede (1980).

Dans une enquête longitudinale, nous devons aussi nous préoccuper de la validité et de la fiabilité des mesures répétées. Comme nous l'avons déjà indiqué, il est fréquent d'observer ce que l'on appelle des « effets de durée de la présence dans l'échantillon », en vertu desquels la proportion de réponses a tendance à évoluer à la hausse ou à la baisse à mesure que la cohorte progresse, simplement à cause du fait d'être mesurée.

Tous ces problèmes accroissent les défis analytiques auxquels font face les chercheurs. Fondamentalement, le but de l'analyse doit être de discerner la réponse réelle (ou le manque de réponse) au changement de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Il s'agit là d'une tâche de taille.

3. Regroupement de données provenant de divers pays

Selon le paradigme classique de l'analyse des données d'enquête (Binder 1983 ; Godambe et Thompson 1986 ; Skinner 1989), nous avons un modèle pour les réponses y avec le paramètre θ, et nous imaginons comment nous estimerions θ si nous possédions les réponses pour l'ensemble de la population grâce à un recensement. Nous utiliserions une équation d'estimation sans biais efficace telle que :

$$\sum_{i=1}^{N} \phi_i(y_i, \theta) = 0,$$

pour définir une *estimation par recensement*. Pour obtenir l'estimation par sondage, nous utilisons une somme pondérée des termes de la fonction d'estimation par sondage :

$$\sum_{i \in s} w_i \phi_i(y_i, \theta) = 0$$

pour obtenir un estimateur approximativement sans biais de la fonction d'estimation par recensement. Nous construisons les poids de sondage de manière à tenir compte du plan d'échantillonnage et de la sous-représentation de certains groupes à cause de la non-réponse et de la non-couverture. L'interprétation habituelle de w_i est le nombre de membres de la population représentée par i. L'utilisation de cette fonction d'estimation par sondage est séduisante, à cause de la réduction probable du biais dû à l'échantillonnage informatif et à la non-réponse ; mais, si les poids sont très variables et que le modèle pour les termes est correct, la deuxième équation donne un moyen inefficace d'estimer θ.

Or, quand nous combinons les données provenant de deux pays pour lesquels les fractions d'échantillonnage sont très différentes, comme dans le cas de l'enquête Irlande/R.-U., les poids pour un pays (R.-U.) seront beaucoup plus grands que ceux pour l'autre pays (Irlande). Si nous appliquons littéralement le paradigme, les données provenant du Royaume-Uni domineront l'analyse. Si le modèle est correct, l'estimation par recensement la plus efficace est la moyenne de y sur l'ensemble des deux pays. Mais alors, l'estimation par sondage correspondante représente une utilisation inefficace de l'échantillon. Ce problème est semblable à celui qui se pose dans les études cas-témoins, qui a été discuté par Scott (2006).

Un moyen de produire de meilleures estimations tout en retenant le paradigme classique consiste à considérer que la valeur du paramètre pour le Royaume-Uni est $\theta - \Delta$, que celle pour l'Irlande est $\theta + \Delta$, et que nous essayons d'estimer θ, la moyenne arithmétique des deux. Un système de fonctions d'estimation par recensement efficace pour θ et Δ équivaut à un système qui se divise en une partie pour chaque pays. Puisque le rééchelonnement des poids dans un pays n'a alors aucun effet sur les estimateurs ponctuels et sur leurs propriétés, la version par sondage pondérée de ce système produit une estimation efficace.

De surcroît, l'analyse qui s'ensuit est approximativement la même que celle que nous obtiendrions suivant le paradigme original si nous avions des tailles d'échantillon égales dans les deux pays et que nous rééchelonnions les poids de manière que leur somme soit égale à la taille de l'échantillon dans chaque pays. Comme l'a souligné Scott (2006), cette façon de rééchelonner les poids est très courante chez les épidémiologistes. Il s'agit, dans un certain sens, d'une application partielle de la méthode du poids q de Pfeffermann et Sverchkov (1999), où le poids égal à l'inverse de la probabilité d'inclusion est divisé par une sorte d'espérance du poids, conditionnellement à une variable explicative (pays).

Pour estimer un paramètre moyen θ, une proposition un peu plus attrayante est de considérer un modèle à effets aléatoires, où $Y = \theta + u + e$, et u est un effet aléatoire de pays, puis d'élaborer un système de fonctions d'estimation par recensement qui est efficace pour l'estimation du paramètre θ. Par exemple, si

$$Y_{1i} = \theta + u_1 + e_{1i} \quad \text{et} \quad Y_{2i} = \theta + u_2 + e_{2i}$$

et si les composantes de la variance correspondant à u et à e sont connues, alors la meilleure combinaison des deux moyennes de pays pour estimer θ est

$$a\bar{Y}_1 + (1-a)\bar{Y}_2,$$

où

$$a = \frac{1}{2}\left\{\frac{\sigma_u^2 + \sigma_e^2/N_2}{\sigma_u^2 + \frac{\sigma_e^2}{2N_1} + \frac{\sigma_e^2}{2N_2}}\right\}.$$

Notons que, si $\sigma_u^2 = 0$, l'estimateur par recensement devient la moyenne de y sur les deux pays combinés. Cependant, si σ_u^2 est dominante, le meilleur estimateur est la moyenne arithmétique des moyennes de pays. Si l'on part d'un échantillon groupé, le paradigme habituel donne la même combinaison convexe d'estimateurs de la moyenne intra-pays fondée sur l'échantillon.

D'une manière plus générale, nous pouvons remplacer θ dans chaque fonction d'estimation par recensement de pays par $\theta + u$, où u désigne de nouveau un effet aléatoire de pays. Alors, la meilleure combinaison des deux fonctions d'estimation par recensement de pays pour θ est

$$c_1 \sum_{i=1}^{N_1} \phi_{1i}(Y_{1i},\ \theta,\ u_1) + c_2 \sum_{i=1}^{N_2} \phi_{2i}(Y_{2i},\ \theta,\ u_2)$$

où $c_1 = [\mathrm{Var}(\sum_{i=1}^{N_2} E(\phi_{2i}|u_2)) + E(\sum_{i=1}^{N_2} \mathrm{Var}(\phi_{2i}|u_2))][\sum_{i=1}^{N_1} E(\partial\phi_{1i}/\partial\theta)]$, et c_2 est défini symétriquement. Dans c_1, si le premier terme entre crochets domine, la fonction d'estimation par sondage correspondante pondère les termes de manière comparable dans les deux échantillons.

Même dans le cas simple d'une moyenne, les paramètres du modèle à effets aléatoires seront inconnus et difficiles à estimer si l'on n'a affaire qu'à deux pays, mais conceptuellement, le modèle semble être utile. Quand il existe plusieurs pays ou régions raisonnablement semblables (par exemple, les sept villes de l'enquête ITC-Chine), des modèles linéaires avec effets aléatoires peuvent être estimés suivant le paradigme habituel, comme il est décrit, par exemple, dans des conditions plus générales par Pfeffermann, Skinner, Holmes, Goldstein et Rasbash (1998).

À titre d'à-côté, l'analyse par EEG des données recueillies en Irlande décrite plus haut était une analyse groupée et tous ses « effets » ont été considérés comme étant fixes. Le modèle est presque « saturé », deux points dans le temps et deux pays représentant les quatre paramètres principaux. On peut voir qu'avec les poids de sondage ordinaires, l'estimation de β et de ρ serait dominée par les données du Royaume-Uni. Cependant, si β et ρ sont connus, comme dans le cas des paramètres θ et Δ dans l'exemple de la moyenne, les équations pour les paramètres principaux se séparent en deux paires, l'une pour α_0 et γ, et l'autre pour $\alpha_0, \alpha_1, \gamma$ et δ, chacune comprenant des poids ne provenant que d'un des deux pays. Donc, l'estimation des paramètres principaux est moins affectée par le rééchelonnement des poids. Si l'estimation de β nous importait aussi, nous pourrions considérer qu'il s'agit de la moyenne d'une variable aléatoire au niveau du pays, ce qui mène naturellement à l'obtention de l'influence appropriée pour chacun des deux échantillons. (En fait, dans notre analyse, nous n'avons pas choisi cette option ; nous avons rééchelonné les poids de manière que leur somme soit égale à la taille de l'échantillon dans chaque pays.)

La discussion qui précède de l'analyse de données groupées repose sur l'hypothèse qu'il existe un paramètre θ dont l'interprétation et la pertinence sont les mêmes dans tous les pays. La plupart des analyses multinationales partent de cette hypothèse. En effet, de Leeuw et Hox (2003) énoncent comme exigence pour une méta-analyse « que toutes les études doivent estimer le même paramètre fixe et qu'il est supposé que toute la variance est égale à la variance d'échantillonnage ». Mais en fait, une question essentielle est celle de savoir s'il est approprié ou non de construire un modèle qui doit s'appliquer simultanément aux données provenant de plusieurs pays. Parfois, il pourrait être plus indiqué de considérer simplement les modèles de pays comme étant distincts, mais parallèles. Par exemple, pour des pays à différents stades de développement, on peut s'attendre à ce que l'introduction de la même augmentation relative du prix réel des cigarettes entraînera une diminution de la consommation de cigarettes ; mais, puisque le modèle linéaire est, au mieux, une approximation locale utile de la relation complexe entre le prix et la consommation, il n'y a aucune raison de supposer que les diminutions seront du même ordre de grandeur ou que les deux estimations par régression mesureront la même quantité.

Voici un autre exemple. L'un des modèles d'intérêt dans le projet ITC est le modèle médiateur de la figure 2, dans lequel il est postulé de quelle façon « remarquer » les étiquettes de mise en garde concernant la santé pourrait influer sur l'intention d'arrêter de fumer.

À la période de référence, la distribution de l'intention d'arrêter varie assez bien selon le pays. Il en est de même des autres variables du modèle. Serait-il raisonnable d'espérer que les relations entre ces variables puissent devenir moins différentes entre les pays ? En fait, il semble que, pour les quatre pays originaux, il en soit ainsi. Même si les fumeurs du Royaume-Uni étaient beaucoup plus susceptibles de déclarer qu'ils n'avaient aucune intention d'arrêter, il n'en reste pas moins que la « préoccupation pour la santé » (prise de conscience des effets nocifs pour la santé déclenchée par les étiquettes) prédisait l'intention d'arrêter et que la prévalence de cette préoccupation augmentait avec le fait de remarquer les étiquettes ((Hammond, Fong, Borland, Cummings, McNeill et Driezen 2007). Donc, il n'est pas déraisonnable d'explorer un modèle tel que celui de la figure 2 pour les données en provenance des quatre pays, regroupées. Indépendamment des problèmes de pondération,

dans la régression du médiateur « préoccupation concernant la santé » sur le « fait de remarquer » les étiquettes de mise en garde concernant la santé, de même que dans la régression de l'intention d'arrêter de fumer sur ces deux variables, nous avons trouvé qu'il était commode de considérer les moyennes de pays comme étant des effets fixes. Par ailleurs, puisque les coefficients de régression estimés pour les pays modélisés séparément varient modérément, il est naturel, dans l'analyse groupée, de conceptualiser ces coefficients comme possédant des composantes de pays aléatoires.

Cette discussion peut se résumer et s'expliquer comme il suit.

- Toute analyse qui consiste à regrouper des données provenant de plusieurs pays devrait être effectuée avec prudence. Pour qu'une telle analyse soit appropriée, la structure du modèle (l'équation de régression et ses variables) doit être correcte pour tous les pays et l'hypothèse que les paramètres sont communs doit être appuyée par la théorie et l'observation. Une estimation de variance robuste qui respecte les plans de sondage des pays sera nécessaire si ces plans sont complexes.
- Si l'ensemble de paramètres d'un modèle groupé peut (par transformation) être séparé en sous-ensembles disjoints correspondant aux pays, l'estimation de ces paramètres n'est pas affectée par de grands écarts entre les fractions d'échantillonnage appliquées dans les divers pays, ni par le rééchelonnement des poids dans les pays.
- Si une moyenne ou un paramètre de régression fixe est considéré comme étant commun aux pays, l'estimation en utilisant comme pondération l'inverse des probabilités d'inclusion sera inefficace si les fractions d'échantillonnage sont très variables.
- L'alternative au simple rééchelonnement des poids consiste à faire de la moyenne ou du paramètre de régression un effet fixe qui varie selon le pays (ce qui aboutit à la séparation en sous-ensembles disjoints, mais augmente le nombre de paramètres et élimine le « caractère commun ») ou à faire de la moyenne ou du paramètre de régression un effet aléatoire, variable selon le pays (ce qui aboutit à une séparation approximative et retient le « caractère commun »).
- Du point de vue conceptuel, il est séduisant de rendre l'ordonnée à l'origine fixe et la pente, aléatoire, puisque le niveau de référence a tendance à varier beaucoup plus selon le pays que la pente. À l'étape de la mise en œuvre, cette approche requiert un nombre suffisant de pays pour que l'estimation des composantes de la variance soit faisable et un petit nombre d'effets aléatoires à intégrer.
- Quand une analyse groupée pose des problèmes, une comparaison moins formelle des résultats des analyses exécutées parallèlement dans les divers pays permet parfois d'accomplir la plupart de ce qu'il est souhaité.

4. Étalonnage des mesures et comparaisons transculturelles

L'autre problème statistique que je souhaite mettre en relief est l'utilisation de modèles de mesure pour essayer d'étalonner les mesures obtenues par divers modes de collecte et de comparer les mesures provenant de diverses cultures. Une approche courante consiste à considérer qu'à l'aide de chaque item du questionnaire, on mesure un concept, comme la « dénormalisation sociale » (perception d'une désapprobation sociétale) et à traiter le concept comme une variable continue η. La distribution de η, sachant les variables explicatives, détermine une distribution des réponses aux items du questionnaire.

Si nous disposons de plusieurs items du même type pour mesurer un concept, un modèle conceptuel pour les mesures continues y pourrait être $Y_{ik} = b_i\eta_k + a_i + e_{ik}$ pour l'item i et le participant k. Ici b_i représente un facteur d'échelle positif pour l'item i, a_i, un changement d'emplacement et e_{ik}, une erreur de mesure normale de moyenne nulle et de variance σ^2_{ei} indépendante de k. Supposons que toutes les e_{ik} sont indépendantes les unes des autres et indépendantes des η_k. (Cela revient en fait à émettre l'hypothèse que η est le seul déterminant latent de Y.) Si nous posons que la distribution de η est $N(0, 1)$, ce que nous pouvons faire si η est normal sans variables explicatives, alors la distribution de Y_{ik} est $N(a_i, b_i^2 + \sigma^2_{ei})$, et, pour un item i unique, les paramètres a_i et $b_i^2 + \sigma^2_{ei}$ peuvent être estimés d'après les données de marge pour un grand nombre de participants. S'il existe au moins deux items ayant la même variance, puisque les covariances des réponses aux items d'un participant sont de la forme $b_{i_1} b_{i_2}$, tous les paramètres peuvent être estimés d'après les données de marge sur de nombreux participants. Sachant les valeurs des paramètres d'item, la valeur de η pour un participant peut être « prédite » d'après la distribution a posteriori de η, sachant les réponses aux items du participant.

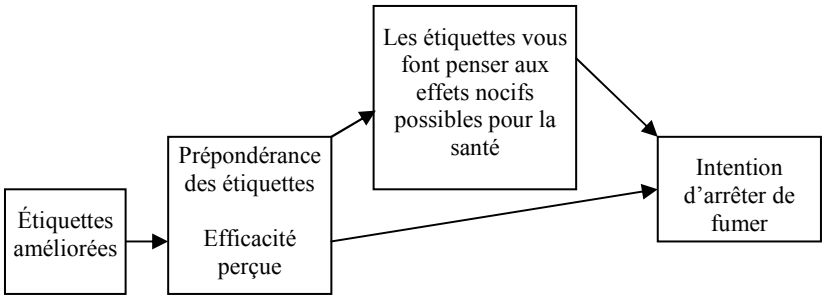


Figure 2 Modèle médiateur des effets des politiques : étiquettes de mise en garde

Si la mesure y est binaire, il est fréquent de choisir un modèle fondé sur la théorie de la réponse à l'item (TRI) $\text{Prob}(Y_{ik} = 1|\eta_k) = H(b_i\eta_k - \gamma_i)$, où H est la loi normale standard ou la fonction de répartition (c.d.f.) logistique. Le paramètre b_i est le « paramètre de discrimination » pour l'item et γ_i est un seuil tel que la probabilité de réponse 1 soit supérieure à 1/2 quand le concept rééchelonné par b_i excède γ_i. La probabilité inconditionnelle que $Y_{ik} = 1$ s'obtient par intégration par rapport à la distribution de η_k, étant donné des variables explicatives fixes pour le participant k. Dans le cas le plus simple, il semble qu'au moins trois items sont nécessaires (avec de nombreux participants) pour que tous les paramètres puissent être estimés, puisqu'ils produiraient sept probabilités conjointes pour l'estimation de six paramètres. De nouveau, sachant les valeurs des paramètres d'item, la valeur de η pour un participant peut être prédite, sachant son ensemble de réponses aux items. Voir, par exemple, Lu, Thomas et Zumbo (2005). Un logiciel standard d'estimation des variables latentes peut être utilisé pour produire ces inférences, ainsi que leurs analogues dans le cas de mesures ordinales.

Commençons par examiner le problème d'étalonnage. Supposons qu'il existe deux modes de collecte des données et que, pour l'item i dans le mode j avec le participant k, nous avons la mesure continue.

$$Y_{ijk} = \beta_j(b_i\eta_k + a_i + e_{ijk}) + \alpha_j + \varepsilon_{ijk}.$$

Ce modèle, dans lequel α_j et β_j ne dépendent pas de l'item i, pourrait convenir pour un ensemble d'items qui sont tous du même type général. Les exemples plausibles ne sont pas nombreux, mais l'un d'eux pourrait être une série de questions de la forme : « Pendant quel pourcentage du temps diriez-vous que vous vous sentez… », où il est demandé au répondant de donner un pourcentage par téléphone, ou de marquer une position sur une ligne dans le cas d'un questionnaire papier.

Si nous posons que les a_i et b_i sont les paramètres des items lorsqu'on utilise le premier mode de collecte des données, nous pouvons fixer $\alpha_1 = 0$ et $\beta_1 = 1$. Si β_2 est plus grand que 1, les réponses ont tendance à être plus variables ou plus extrêmes sous le deuxième mode de collecte. Si α_2 est plus grand que 0, les répondants ont tendance à donner une réponse plus élevée sous le deuxième mode de collecte que sous le premier. Notons que les échantillons pour les deux modes de collecte sont constitués de participants différents. Si nous pouvons émettre l'hypothèse que la distribution de η est la même pour les deux échantillons (une hypothèque qui, en fait, requiert une randomisation par rapport au mode), nous obtenons la distribution de Y_{i1k} comme auparavant, $N(a_i,\ b_i^2 + \sigma_{ei}^2 + \sigma_\varepsilon^2)$, tandis que la distribution de Y_{i2k} est

$$N(\beta_2\, a_i + \alpha_2,\ \beta_2^2 b_i^2 + \beta_2^2 \sigma_{ei}^2 + \sigma_\varepsilon^2).$$

Si $\sigma_\varepsilon^2 = 0$, sachant les données sur un item i dans les deux modes, nous pouvons estimer α_2 et β_2, en supposant que β_2 est positif. Si $\sigma_\varepsilon^2 > 0$, les paramètres α_2, β_2 et σ_ε^2 peuvent être estimés, à condition que l'on dispose d'au moins deux items – du même type, mais dont les valeurs de a et b diffèrent.

Ces considérations peuvent être étendues au cas plus habituel des items avec réponses ordinales en imaginant une probabilité de réponse ordinale devant être déterminée par une réponse continue sous-jacente. Pour des données binaires, nous poserions le plus simplement

$$P(Y_{ijk} = 1|\eta_k) = H(\beta_j(b_i\eta_k + a_i) + \alpha_j),$$

avec $\alpha_1 = 0$ et $\beta_1 = 1$. Si la distribution de η_k est la même pour les deux modes, alors, avec des données provenant de nombreux participants et trois items, nous pouvons déterminer tous les paramètres. L'ajout d'une variable explicative réduirait le nombre d'items requis.

L'hypothèse que la distribution de η est la même pour les échantillons correspondants aux deux modes est cruciale pour ce type d'étalonnage, et il est difficile de garantir qu'elle sera vérifiée. Elle l'est si nous avons des échantillons probabilistes superposés pour les deux modes dans une seule enquête ; alors, en principe, nous pouvons imaginer une mise en correspondance des réponses d'un mode à

l'autre, par la voie des valeurs estimées de α_2 et β_2. Nous ne devons pas estimer les concepts proprement dits pour cela. Plus rigoureusement, nous pouvons inclure α_2 et β_2 comme paramètres dans un modèle pour toutes les réponses à un ensemble d'items similaires.

Dans certains pays développés, les bases de sondage pour les ménages et les particuliers semblent évoluer vers des registres d'adresses et des listes de personnes. Cependant, même s'il existe une base de sondage commune pour (disons) les enquêtes par téléphone et en ligne, il est difficile de randomiser les répondants par rapport aux modes de collecte des données. La façon dont la non-réponse dépend des variables démographiques pourrait fort bien différer selon le mode. En outre, la nécessité de maximiser les taux de réponse oblige souvent à permettre aux répondants de choisir. En principe, nous pourrions imaginer que la distribution de η peut être déplacée ou inclinée en fonction de la « propension » à choisir un mode ou l'autre. Après avoir modélisé cette propension en fonction des variables explicatives et introduit un ou deux paramètres pour la dépendance de la distribution de η à l'égard de la propension, nous pourrions estimer les paramètres d'item a_i et b_i d'après les réponses obtenues par le premier mode de collecte des données. L'estimation des paramètres de mise en correspondance α et β s'ensuivrait de la même manière qu'auparavant.

Dans d'autres circonstances, nous pourrions utiliser les deux modes de collecte des données dans différents groupes de population. Dans ce cas, l'effet de mode fait partie de l'effet de groupe ; il ne peut pas être distingué d'une différence sous-jacente dans la distribution du concept.

Le problème de la comparaison des mesures entre différentes cultures ou entre d'autres groupes n'est pas le même que celui de l'étalonnage, puisque la randomisation des participants par rapport aux groupes, afin que la distribution du concept demeure constante, est hors de question. On pense généralement que, pour comparer la moyenne d'un concept d'un groupe à l'autre, les items mesurés doivent avoir la même relation avec le concept dans les deux groupes. Quand il existe plusieurs concepts, la comparaison de la relation entre les concepts d'un groupe à l'autre requiert une sorte d'« invariance de mesure » ou d'équivalence pour tous les items concernés. La littérature sur les comparaisons et les mesures transculturelles est abondante. Par exemple, Johnson (1998) énumère, pour l'équivalence transculturelle, 52 termes qui ont été introduits par les auteurs dans diverses disciplines.

Le modèle d'analyse factorielle confirmatoire multigroupes est utile pour les items dont la mesure est continue et prend la forme :

$$Y_k^g = \tau^g + \Lambda^g \eta_k^g + e_k^g$$

où Y_k^g est le vecteur des réponses observées aux items pour le répondant k dans le groupe g, Λ^g est une matrice des pentes ou des « saturations factorielles », le vecteur d'ordonnées à l'origine τ^g indique la valeur attendue de Y_k^g quand $\eta_k^g = 0$ et e_k^g est une erreur de mesure de moyenne 0. Alors, $E(Y_k^g) = \tau^g + \Lambda^g \kappa^g$, où κ^g est la moyenne du concept η dans le groupe g. La matrice de variance-covariance entre les valeurs observées y_k^g peut être exprimée sous la forme $V(Y_k^g) = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g$, où Φ^g est la matrice de covariance des concepts latents et Θ^g est la matrice diagonale des variances de l'erreur de mesure. À cet égard, voir de Jong, Steenkamp et Fox (2007), Davidov (2008) et les références qu'ils fournissent.

La version TRI du modèle peut être définie simplement. En utilisant la même notation pour les paramètres, dans le cas des items binaires, nous avons

$$P(Y_k^g = 1 \mid \eta_k^g) = H(\tau^g + \Lambda^g \eta_k^g),$$

et il existe une extension naturelle au cas ordinal.

Les paramètres du modèle ne peuvent pas être identifiés, à moins d'imposer certaines contraintes. Dans le modèle d'analyse factorielle confirmatoire multigroupes, de nombreux auteurs postulent un item « marqueur » pour chaque concept, avec une saturation factorielle de 1 et une ordonnée à l'origine de 0 pour tous les groupes, de sorte que la moyenne du concept est déterminée dans chaque groupe. Cette hypothèse est très forte. Alternativement, nous pourrions choisir les unités pour les concepts de sorte qu'elles soient marginalement N(0, 1) dans le groupe 1. Les paramètres des items (avec un nombre suffisamment grand d'items) sont donc identifiés pour le groupe 1. Si nous supposons que les variances et les relations du diagramme des trajectoires restent vraies dans le groupe 2, alors nous pouvons vérifier si les paramètres d'item restent également les mêmes et, dans la négative, essayer de reconcevoir l'ensemble d'items afin d'en produire un s'approchant davantage de l'invariance de mesure. Par ailleurs, si les paramètres d'item sont contraints de demeurer les mêmes, nous pouvons vérifier si la distribution conjointe sous-jacente des concepts est également la même. Cependant, le rejet formel de l'hypothèse nulle est difficile à interpréter. S'inspirant de Rensvold et Cheung (1998), Barrera Ceballos (2007) a procédé à ce genre d'analyse multigroupes pour les données de l'enquête ITC-Mexique et de l'enquête ITC-Uruguay, en remplaçant « préoccupation concernant la santé » dans le modèle de la figure 2 par « dénormalisation sociale », c'est-à-dire la mesure dans laquelle le répondant a l'impression que la société désapprouve le tabagisme. (Les deux autres concepts sont la prépondérance des étiquettes de mise en garde et l'intention d'arrêter.) Les relations semblent étonnamment différentes dans les deux pays sous les contraintes d'invariance des items de mesure, résultat qui

pourrait être dû à des différences sociétale réelles ou à une correspondance imparfaite entre les items proprement dits (c'est-à-dire échec des contraintes). Il faut reconnaître que, ne contenant qu'un très petit nombre de concepts possédant des items multiples, le questionnaire de l'enquête ITC n'était pas conçu pour ce genre d'analyse.

En fin de compte, les relations entre les concepts sont de la plus haute importance, de même que la question de savoir si ces relations peuvent être considérées comme étant semblables, quoique pas forcément identiques, d'un groupe à l'autre. Il en est ainsi, que les distributions marginales des concepts soient les mêmes ou que les items mesurés possèdent les mêmes paramètres d'un lieu à l'autre ou d'un mode à l'autre. Intuitivement, les deux catégories de contraintes mentionnées au paragraphe précédent semblent trop fortes. Une approche hiérarchique proposée par De Jong et coll. (2007) offre une solution.

Si l'item i comprend C options de réponse ordonnées, nous pouvons écrire

$$P(Y_{ik}^g = c \mid \eta_k^g, b_i^g, \gamma_{i,c}^g, \gamma_{i,c-1}^g) = H(b_i^g \eta_k^g - \gamma_{i,c-1}^g) - H(b_i^g \eta_k^g - \gamma_{i,c}^g),$$

$c = 1, \ldots, C$. Ici, les saturations factorielles sont remplacées par les paramètres de discrimination b, et les ordonnées à l'origine sont remplacées par les seuils γ. Au lieu d'insister sur le fait que ces paramètres sont indépendants de l'étiquette de groupe avant le traitement, l'approche consiste à les modéliser avec des effets aléatoires propres au groupe :

$$\gamma_{i,c}^g = \gamma_{i,c} + e_{i,c}^g,\ e_{i,c}^g \sim N(0, \sigma_{\gamma_i}^2),$$
$$b_i^g = b_i + r_i^g,\ r_i^g \sim N(0, \sigma_{b_i}^2).$$

L'hétérogénéité de la variable latente est modélisée par une structure hiérarchique :

$$\eta_k^g = \kappa^g + v_k^g,\ v_k^g \sim N(0, \sigma_g^2),$$
$$\kappa^g \sim N(\kappa, \xi^2).$$

Si le nombre d'items est suffisant, un tel modèle peut être estimé et ajusté en utilisant des méthodes Monte Carlo à chaîne de Markov. Les tests d'invariance de l'analyse multigroupes peuvent encore être exécutés dans ce cadre de travail.

5. Discussion et conclusion

De nouveau, dans le contexte du projet ITC, le but de l'analyse doit être de discerner la réponse réelle (ou le manque de réponse) au changement de politiques et de la séparer des effets du mode de collecte des données, de la non-réponse différentielle, des événements extérieurs, de la durée de la présence dans l'échantillon, de la culture et de la langue. Il n'est pas toujours nécessaire de faire la distinction entre tous les facteurs confusionnels, mais il est important de leur permettre de contribuer au modèle. Dans le présent article, nous n'avons pas abordé la question des événements externes, qui peuvent être modélisés de manière évidente s'ils sont reconnus. Nous n'avons pas discuté en détail de la modélisation des effets d'attrition et de durée de la présence dans l'échantillon, mais en principe, chacun de ces effets peut être considéré comme faisant partie de l'ensemble. Ceux qui sont retenus d'une vague à l'autre d'une enquête pourraient être considérés comme une sorte de groupe culturel. Par ailleurs, les effets de durée de la présence dans l'échantillon constituent une forme particulière d'échec de l'invariance de la mesure, au cours du temps, plutôt que d'un groupe à l'autre. Une analyse complète tiendrait compte de ces effets, ainsi que d'autres effets de la culture, de la langue et du mode de collecte de données.

Il n'est absolument pas certain que les effets des politiques puissent systématiquement être identifiés dans un modèle complet. Mais les chances augmentent si le plan de sondage tient compte des comparaisons de données longitudinales entre pays et de la répétition qui provient de l'observation de cohortes qui ont des points de départ différents.

Un thème commun aux deux sections précédentes est l'introduction d'effets aléatoires en tant qu'outil. Le procédé consistant à introduire des effets aléatoires pour les pays et les groupes dans les paramètres clés est naturel et (pour les échantillons de grands groupes) conceptuellement compatible avec l'analyse classique des données d'enquête fondées sur les fonctions d'estimation pondérées. Il existe certains obstacles à l'application pratique, dus aux limites d'identifiabilité et d'estimabilité, ainsi qu'au calcul des fonctions de vraisemblance si l'on envisage plus qu'un ou deux effets aléatoires. Parallèlement, étant donné la disponibilité croissante de méthodes numériques pour traiter ce genre de modèle, la poursuite des travaux de recherche en vue de les adapter à des enquêtes internationales complexes devrait être très fructueuse.

Remerciement

Ces travaux sont financés partiellement par une subvention du Conseil de recherche en sciences naturelles et en génie du Canada. Le projet ITC est financé par des subventions du National Cancer Institute des États-Unis (P50 CA11236), du Roswell Park Transdisciplinary Tobacco Use Research Center et des Instituts de recherche en santé du Canada (57897). L'auteure remercie un examinateur anonyme de ses commentaires très constructifs.

Bibliographie

Barrera Ceballos, J.A. (2007). Cross-national comparison of the impact of cigarette warning labels and social denormalization on intention to quit from the International Tobacco Control Survey. Research paper. Statistics and Actuarial Science, University of Waterloo.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Bronner, K., et Kuijlen, T. (2007). The live or digital interviewer: A comparison between CASI, CAPI and CATI with respect to differences in response behaviour. *International Journal of Market Research*, 49, 167-190.

Cleland, J., et Verma, V. (1989). The World Fertility Survey: An appraisal of methodology. *Journal of the American Statistical Association*, 84, 756-767.

Cook, T., et Campbell, D. (1979). *Quasi-Experimentation*. Chicago : Rand McNally.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods*, 2, 33-46.

De Jong, M.G., Steenkamp, J.-B.E.M. et Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical irt model. *Journal of Consumer Research*, 34, 260-278.

de Leeuw, E.D., et Hox, J. (2003). The use of meta-analysis in cross-national studies. Dans *Cross-Cultural Survey Methods* (Éds., J.A. Harkness, F.J.R. Van de Vijver et P. Ph. Mohler). Hoboken, NJ : Wiley, 329-346.

ESS (2008). European Social Survey. http://www.europeansocialsurvey.org/.

Fong, G.T., Cummings, K.M., Borland, R., Hastings, G., Hyland, A., Giovino, G.A., Hammond, D. et Thompson, M.E. (2006). The conceptual framework of the International Tobacco Control Policy Evaluation Project. *Tobacco Control*, 15(Supp 3) : iii3-iii11.

Fong, G.T., Hyland, A., Borland, R., Hammond, D., Hastings, G., McNeill, A., Anderson, S., Cummings, K.M., Allwright, S., Mulcahy, M., Howell, F., Clancy, L., Thompson, M.E., Connolly, G. et Driezen, P. (2006). Reductions in tobacco smoke pollution and increases in support for smoke-free public places following the implementation of comprehensive smoke-free workplace legislation in the Republic of Ireland: Findings from the ITC Ireland/UK Survey. *Tobacco Control*, 15(Supp. 3) : iii51-iii58.

Godambe, V.P., et Thompson, M.E. (1986). Parameters of superpopulation and survey populations: Their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.

Hammond, D., Fong, G.T., Borland, R., Cummings, K.M., McNeill, A. et Driezen, P. (2007). Text and graphic warnings on cigarette packages: Findings from the International Tobacco Control Four Country Study. *American Journal of Preventive Medicine*, 32, 210-217.

Hofstede, G. (1980). *Culture's Consequences. International Differences in Work-Related Values.* Beverly Hills, CA : Sage.

Hyland, A., Hassan, L., Higbee, C., Fong, G.T., Borland, R., Cummings, K.M., Thompson, M., Boudreau, C. et Hastings, G. (2008). The impact of smokefree legislation in Scotland: Results from the Scottish International Tobacco Control Policy Evaluation Project. Travaux en cours.

ITC (2008). Tobacco Labelling Resource Centre. http://www.igloo.org/tobacco_labelling. Accessible le 24 avril 2008.

Johnson, T.P., et Van de Vijver, F.J.R. (2003). Social desirability in cross-cultural research. Dans *Cross-Cultural Survey Methods* (Éds., J.A. Harkness, F.J.R. Van de Vijver et P. Ph. Mohler). Hoboken, NJ : Wiley, 195-204.

Johnson, T.P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. Dans *Cross-Cultural Survey Equivalence* (Éd., J.A. Harkness). *ZUMA-Nachrichten Spezial 3*. Mannheim : ZUMA, 1-40.

Lohr, S.L., et Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lu, I.R.R., Thomas, D.R. et Zumbo, B.D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling*, 12, 263-277.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. et Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Séries B, 60, 23-40.

Pfeffermann, D., et Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, Séries B, 61, 166-186.

Rensvold, R.B., et Cheung, G.W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, 58, 1017-1034.

Scott, A. (2006). Études cas-témoins basées sur la population. *Techniques d'enquête*, 32, 137-147.

Rosenbaum, P.R., et Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-542.

Ryder, A.G., Yang, J., Zhu, X., Yao, S., Yi, J., Heine, S.J. et Bagby, R.M. (2008). The cultural shaping of depression: Somatic symptoms in China, psychological symptoms in North America? *Journal of Abnormal Psychology*, 117, 300-313.

Skinner, C. (1989). Introduction to Part A. Dans *Analysis of Complex Surveys* (Éds., C. Skinner, D. Holt et T.M.F. Smith), Chichester : Wiley. 2.

Thompson, M.E., Boudreau, C. et Driezen, P. (2005). Incorporating time-in-sample in longitudinal survey models. *Recueil : Symposium 2005, Défis méthodologiques reliés aux besoins futures d'information.* Session 12 : Défis pour l'emploi des données provenant d'enquêtes longitudinales. Statistique Canada.

Thompson, M.E., Fong, G.T., Hammond, D., Boudreau, C., Dreizen, P., Hyland, A., Borland, R., Cummings, K.M., Hastings, G.B., Siahpush, M., Mackintosh, A.M. et Laux, F.L. (2006). Methods of the International Tobacco Control (ITC) Four Country Survey. *Tobacco Control*, 15(Supp 3) : iii12-iii18.

Thrasher, J., Quah, A., Borland, R., Awang, R., Sirirassamee, B., Boado, M., Miller, K., Watts, A. et Dorantes, A. (2008). Ensuring valid cross-cultural comparisons in survey research on tobacco: Development, implementation, and results from a transnational cognitive interviewing study. Travaux en cours.

Verma, V., Scott, C. et O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, Séries A, 143, 431-473.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

WHO (2008). *WHO Report on the Global Tobacco Epidemic*, *2008*. http://www.who.int/tobacco/mpower/mpower_report_full_2008.pdf Accessible le 14 avril 2008.

Wichers, B., et Zengerink, E. (2006). It's the culture, stupid! A cross-cultural comparison of data collection methods. *Panel Research 2006*, *Part 4/The respondent - Cross cultural insights*, ESOMAR.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 159-168
Statistique Canada, N° 12-001-X au catalogue

Utilisation du codage du comportement pour évaluer l'efficacité de l'interview avec rétroinformation

Joanne Pascale et Alice McGee [1]

Résumé

L'interview avec rétroinformation (IRI) est utilisée dans de nombreuses enquêtes longitudinales pour « reporter » les données d'une vague à la suivante. Bien qu'il s'agisse d'une technique prometteuse qui a permis d'améliorer la qualité des données à certains égards, on en sait assez peu sur la façon dont elle est effectivement appliquée sur le terrain. La présente étude a pour but d'aborder cette question par la voie du codage du comportement. Divers styles d'IRI ont été employés dans l'English Longitudinal Study of Ageing (ELSA) de janvier 2006 et les interviews pilotes sur le terrain ont été enregistrées. Les enregistrements ont été analysés afin de déterminer si les questions (particulièrement les éléments d'IRI) étaient administrées convenablement et d'explorer la réaction des répondants à l'égard des données reportées. Un des objectifs était de déterminer si les répondants confirmaient ou contestaient l'information déclarée antérieurement, si les données de la vague précédente jouaient un rôle quand les répondants fournissaient les réponses à la vague courante et la façon dont toute discordance était négociée par l'intervieweur et le répondant. Un autre objectif était d'examiner l'efficacité de divers styles d'IRI. Par exemple, dans certains cas, les données de la vague précédente ont été reportées et on a demandé aux répondants de les confirmer explicitement ; dans d'autres, les données antérieures ont été lues et ont a demandé aux répondants si la situation était encore la même. Les résultats révèlent divers niveaux de conformité en ce qui a trait à la lecture initiale de la question et donnent à penser que certains styles d'IRI pourraient être plus efficaces que d'autres.

Mots clés : Interview avec rétroinformation ; enquêtes longitudinales ; enquêtes par panel ; codage du comportement.

1. Introduction

Ces dernières années, l'intérêt pour l'« interview avec rétroinformation » (ou IRI) et son emploi dans les enquêtes longitudinales se sont accrus. L'IRI (également appelée interview « avec données déclarées antérieurement ») est une technique qui consiste à reporter les données recueillies durant une vague à la vague suivante afin de personnaliser l'énoncé des questions et leur enchaînement. Par exemple, si un répondant a déclaré à la première vague travailler pour l'employeur X, la question d'IRI posée à la deuxième vague sera : « La dernière fois, vous avez dit que vous travailliez pour l'employeur X. Travaillez-vous encore pour l'employeur X ? » En revanche, dans le cas d'une méthode d'interview « indépendante » (c'est-à-dire sans rétroinformation), à la deuxième vague, on demanderait simplement au répondant, « en repartant à zéro », s'il travaille et quel est le nom de l'employeur. Une application connexe de l'IRI consiste à enchaîner les questions de manière à épargner au répondant la tâche de répondre à des questions détaillées si les circonstances n'ont pas changé d'une vague à l'autre. Par exemple, un ensemble détaillé de questions sur l'employeur X sera posé à la première vague (tel que l'industrie, le nombre d'employés, *etc*.), puis à la deuxième vague, si le répondant déclare qu'il travaille encore pour le même employeur, la collecte des renseignements détaillés ne doit pas se faire une deuxième fois.

La prolifération des enquêtes automatisées a favorisé le recours à l'IRI, étant donné que la technique peut être fastidieuse et difficile à appliquer à l'aide d'un questionnaire papier-crayon. La possibilité d'améliorer de plusieurs façons la qualité des données est un autre facteur qui accroît l'intérêt pour cette méthode. En général, l'IRI rend l'interview plus harmonieuse, plus intelligente et plus efficace parce qu'on rappelle aux répondants leurs réponses antérieures et qu'on leur permet tout simplement d'indiquer s'il y a eu des changements depuis. Les données d'études rigoureuses « prouvant » ce potentiel commencent à s'accumuler. Par exemple, il existe des preuves convaincantes que l'IRI réduit les fausses variations, particulièrement celles des caractéristiques de l'emploi (Polivka et Rothgeb 1993 ; Jäckle et Lynn 2004). Une autre source d'erreur de mesure qui affecte régulièrement les enquêtes par panel est le « biais de lisière ».

La « lisière » d'une enquête par panel est le point où une vague est jointe à la suivante. Par exemple, dans une enquête par panel comportant des vagues annuelles, la lisière se situe entre décembre d'une année et janvier de l'année suivante. Le biais de lisière a lieu quand un plus grand nombre de transitions (par exemple de l'emploi au chômage) sont observées de décembre à janvier que pendant toute paire de mois non lisière (par exemple de février à mars, d'avril à mai). Il existe des preuves solides que l'IRI réduit significativement (mais n'élimine pas) ce biais de lisière (Moore, Bates, Pascale et Okon 2006). En ce qui

1. Joanne Pascale, US Census Bureau, 4600 Silver Hill Road, Washington DC 20233. Courriel : joanne.pascale@census.gov ; Alice McGee, anciennement avec le National Centre for Social Research.

concerne l'accueil réservé à l'IRI, il existe des preuves qualitatives que les répondants la veulent ou l'attendent (Pascale et Mayer 2004). Durant l'été 2006, une grande conférence a été organisée pour évaluer l'« état de la recherche » sur les études longitudinales et plusieurs communications ont démontré les avantages particuliers de l'IRI. Un recueil révisé de monographies comprenant certaines communications sera publié par John Wiley and Sons en 2008 (http://www.iser.essex.ac.uk/ulsc/ mols2006/).

Toutefois, jusqu'à présent, la littérature semble ne pas fournir de données sur la façon dont l'IRI est effectivement mise en œuvre sur le terrain. Le propos de la présente étude est de combler cette lacune. En particulier, nous recourons au codage du comportement pour déterminer si les intervieweurs lisent les questions telles qu'elles sont rédigées, en nous concentrant surtout sur les mots et les phrases qui font partie de la rétroinformation dans les questions, et nous examinons les réactions des répondants aux phrases reportées − c'est-à-dire s'ils confirment ou contestent les données recueillies antérieurement, et si la fourniture de cette information semble faciliter ou gêner la déclaration. Enfin, nous déterminons si ces comportements varient selon le « style » d'IRI, c'est-à-dire le moyen particulier utilisé pour redonner aux répondants l'information recueillie antérieurement. L'étude s'appuie sur l'English Longitudinal Study of Ageing (ELSA) exécutée par le National Centre for Social Research (ou « NatCen ») en collaboration avec le University College London et l'Institute of Fiscal Studies.

2. Méthodes

2.1 ELSA : l'instrument d'enquête

ELSA est une étude sur les personnes de 50 ans et plus et leurs partenaires plus jeunes. Elle a pour but d'explorer la dynamique de la santé et de l'incapacité, la structure familiale, la participation à des programmes publics, les circonstances économiques et la retraite. L'échantillon d'ELSA a été sélectionné parmi les ménages qui avaient participé antérieurement à cinq vagues annuelles de la Health Survey for England (HSE) entre 1998 et 2003. La première vague d'ELSA a été exécutée en 2002 auprès de 12 100 répondants et des interviews de suivi ont été menées tous les deux ans pour mesurer les changements d'état de santé et de circonstances socioéconomiques. L'interview avec rétroinformation a été intégrée dans le questionnaire de la deuxième vague, mais à cause de contraintes budgétaires et de temps, l'évaluation avant sa mise en œuvre a été limitée. L'analyse des données de la deuxième vague a suscité certaines préoccupations quant à l'effet de l'IRI. Par exemple, environ 20 % des répondants qui avaient déclaré faire de l'hypertension à la première vague ont dit qu'ils n'en faisaient plus à la deuxième vague. En partie à cause de ce résultat, le présent projet de recherche a été entrepris en vue d'évaluer d'une manière générale la mise en œuvre des techniques d'IRI sur le terrain. Le codage du comportement est la méthode qui a été choisie pour évaluer soigneusement les interactions intervieweur-répondant et pour déterminer dans quelle mesure les questions sont posées telles qu'elles sont rédigées.

2.2 Interview et enregistrement sur le terrain

La phase pilote de la troisième vague d'ELSA a été réalisée sur une période de quatre semaines en janvier 2006. En tout, 17 intervieweurs de terrain de NatCen venant de diverses régions du Royaume-Uni ont réalisé 123 interviews sur place. La grande majorité de ces interviews (106) ont été menées auprès de personnes qui avaient été interrogées à la vague précédente d'ELSA, tandis que 17 ont été effectuées auprès de membres d'un échantillon de remise à niveau qui n'avaient pas encore participé à ELSA, mais qui avaient été interviewés dans le cadre de la HSE. La plupart des analyses présentées ici ont trait aux personnes qui avaient été interviewées durant la vague précédente d'ELSA. Toutefois, pour deux items du module des caractéristiques démographiques − RÉSIDE (savoir si un membre du ménage réside encore à la résidence) et DDN (date de naissance du membre du ménage) − les répondants de la vague antérieure d'ELSA ainsi que les membres de l'échantillon de remise à niveau ont été inclus dans l'analyse, puisque ces questions incluaient des données reportées provenant des interviews de la HSE. Autrement dit, ces deux items comprenaient une IRI, même pour les cas de remise à niveau qui n'avaient pas été interviewés à la vague précédente d'ELSA. Les interviews ont été effectuées sur place en utilisant un questionnaire papier avec assistance par ordinateur (IPAO). Les questions portaient sur plusieurs sujets, dont les caractéristiques démographiques du ménage et des individus, l'état de santé, le revenu et les avoirs. Les interviews ont été enregistrées à l'aide d'une application logicielle d'interview enregistrée assistée par ordinateur (IEAO), qui permet d'enregistrer les intervieweurs de terrain directement sur leur ordinateur portable sous forme de fichier sonore numérique. Une question de consentement demandant aux répondants d'autoriser l'enregistrement de l'interview a été insérée au début du questionnaire et l'enregistreur a été éteint quand le répondant ne donnait pas son consentement. En outre, certains fichiers sonores n'ont pas pu être codés parce qu'ils étaient corrompus. Parmi les 123 interviews sur place, 104 ont été codées et, parmi les 106 interviews de la vague précédente d'ELSA, un total de 87 enregistrements individuels ont été codés. Dans les deux cas, la majorité des

pertes d'interviews était due au non-consentement (par opposition à la corruption des fichiers sonores).

2.3 Énoncé des questions de l'interview avec rétroinformation

L'interview avec rétroinformation a été intégrée dans trois modules thématiques de l'instrument, à savoir les caractéristiques démographiques, les problèmes de santé et la propriété d'un véhicule (voir figure 1). Le module sur les problèmes de santé comportait trois grandes catégories de maladies, à savoir les maladies des yeux, les maladies cardiovasculaires (MCV) et les maladies chroniques. Dans chacune de ces grandes catégories, des questions ont été posées sur de multiples maladies particulières. Par exemple, la catégorie des maladies des yeux en comptait quatre (dont le glaucome et la cataracte). Les items 4 et 5 ont été répétés pour chaque maladie ou affection déclarée par le répondant à la vague précédente.

Cinq styles distincts d'IRI ont été utilisés dans ces trois modules thématiques, mais comme nous l'avons mentionné plus haut, aucune étude particulière n'a orienté ces décisions de conception de l'étude. Comme l'indique la figure 1, un style d'IRI légèrement différent a été employé pour chacun des six items. Les deux premiers items du module sur les caractéristiques démographiques (RÉSIDE, DDN) fournissaient des données déclarées antérieurement, mais ne mentionnaient pas explicitement que ces données avaient été recueillies lors de l'interview précédente. Les données antérieures étaient simplement présentées au répondant et il lui était demandé de les vérifier. Le troisième item (ENFANT) indiquait explicitement que les données avaient été recueillies la dernière fois et demandait au répondant si l'information était correcte. Contrairement aux questions sur les caractéristiques démographiques, celles sur la santé ont été réparties en deux items distincts apparaissant sur deux écrans différents. Le premier (PRÉCÉDANTE-YEUX) était simplement un énoncé informant le répondant qu'une maladie particulière avait été déclarée durant l'interview précédente. Cet item devait être lu comme un énoncé, il n'était pas demandé au répondant d'y répondre et aucune réponse n'était attendue de lui ; l'intervieweur devait au contraire appuyer sur la touche « enter » afin de faire apparaître la deuxième partie de la série. Le deuxième item (ENCORE-YEUX) demandait si le répondant avait encore la maladie ou l'affection. Enfin, pour la propriété d'un véhicule, la routine était semblable dans une certaine mesure à celle pour les questions sur les problèmes de santé ; pour commencer, un énoncé a été lu pour informer le répondant de ce qu'il avait déclaré la dernière fois, puis une question a été posée pour déterminer si la situation déclarée existait encore (c'est-à-dire : Possédez-vous encore le véhicule ?). La différence tient au fait que, pour l'item sur le véhicule, l'énoncé sur la situation passée et la question (« possède encore ») ont été regroupés en un seul item, tandis que dans le module de la santé, deux items distincts ont été utilisés − l'énoncé, puis la question « encore ? ».

A. *Caractéristiques démographiques*
   1. RÉSIDE : NOM réside-t-il/elle encore ici ?
   2. DDN : Puis-je simplement vérifier, est-ce que la date de naissance de NOM est [insérer la date de naissance (DDN)] ?
   3. ENFANT : Nos dossiers indiquent qu'au moment de la dernière interview, vous aviez un enfant nommé NOM, dont la date de naissance est [DDN]. Ces détails sont-ils corrects ?

B. *Problèmes de santé*
   4. PRÉCÉDENTE – YEUX/MCV/CHRON : Nos dossiers indiquent qu'au moment de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [insérer MALADIE DES YEUX/MCV/MALADIE CHRONIQUE]. [appuyez sur la touche enter]
   5. ENCORE-YEUX/MCV/CHRON : Avez-vous encore une [insérer MALADIE DES YEUX/MCV/ MALADIE CHRONIQUE] ?

C. *Propriété d'un véhicule*
   6. VÉHICULE : La dernière fois que nous vous avons vu(e), vous nous avez dit que vous étiez le principal utilisateur d'un(e) [MARQUE DU VÉHICULE], avec une immatriculation [LETTRE]. Possédez-vous encore ce véhicule ?

Figure 1 Énoncé des questions des items s'appuyant sur l'interview avec rétroinformation

2.4 Codage du comportement

Afin d'établir le cadre de codage du comportement, nous avons commencé par écouter plusieurs enregistrements pour nous faire une idée générale du cheminement de l'interview, de la fréquence et de la nature des comportements non standards de l'intervieweur et des réactions des répondants aux questions. Nous avons déterminé que « le premier échange » − la lecture initiale de la question par l'intervieweur et la première énonciation du répondant en réponse à cette question − était suffisamment riche pour l'analyse et nous avons donc établi un cadre de codage pour ne saisir que ces comportements, ainsi qu'un résultat final. Dans le contexte de ces trois comportements (lecture initiale de la question par l'intervieweur, réponse initiale du répondant et résultat), nous sommes partis d'un cadre de codage relativement standard que nous avons adapté en fonction du contenu des enregistrements et de notre souhait particulier d'en apprendre sur le fonctionnement des phrases reportées intégrées dans les questions (voir la figure 2). Pour le comportement de l'intervieweur, nous avons utilisé trois grandes catégories de codes, à savoir 1) la question a été lue telle que rédigée ou avec un changement mineur seulement qui n'a

pas modifié la signification de la question, 2) la question a été lue avec un « changement majeur » qui a modifié ou pourrait modifier la signification de la question et 3) la question a été omise. Dans la catégorie du code de changement majeur, nous avons établi deux codes propres à l'IRI. Dans les enregistrements, il était assez fréquent d'entendre les intervieweurs transformer un énoncé en une question. Par exemple, dans le module sur la santé, l'énoncé : « Nos enregistrements indiquent que la dernière fois, vous avez déclaré l'affection X » devenait une question parce que les intervieweurs ajoutaient souvent « Est-ce correct ? » Ou, dans certains cas, les intervieweurs utilisaient une intonation et une pause qui transformaient l'énoncé en une question − par exemple « Nos dossiers indiquent que la dernière fois vous avez déclaré l'affection X ? » suivi par une pause, en attente d'une réponse du répondant. Il convient de souligner que, comme les codeurs travaillaient directement d'après les enregistrements au lieu des transcriptions, ils étaient capables de porter un jugement quant à l'utilisation d'une intonation évoquant une question ou un énoncé. Dans d'autres cas, une question devenait un énoncé. Par exemple, dans le module sur les caractéristiques démographiques, la question « NOM réside-t-il/elle encore ici ? » a été modifiée en « Et NOM réside encore ici. » (sans intonation indiquant un point d'interrogation). Comme ces problèmes étaient ceux observés le plus fréquemment, nous avons créé des codes spéciaux pour eux.

En tout, nous avons utilisé sept codeurs, provenant des sections des méthodes d'enquête et des opérations. Les chercheurs leur ont donné une formation d'une demi-journée qui a duré quatre heures. Cette formation portait sur les concepts de base du codage du comportement, ainsi que sur les codes particuliers à l'étude et sur la façon de les appliquer. La majorité de la séance a été consacrée au codage d'exemples hypothétiques d'interactions répondant-intervieweur, suivi d'une discussion et d'une comparaison des jugements individuels et de la justification de ces jugements pour tenter d'uniformiser l'application des codes entre les codeurs. Cependant, aucune mesure formelle de fiabilité n'a été mise en œuvre.

Les codes de répondant étaient relativement standard, de nouveau à l'exception de ceux propres à l'IRI. Un code « adéquate » signifiait que l'énonciation initiale du répondant rentrait dans l'une des catégories de réponse. Nous avons adapté ce code pour saisir le fait que le répondant confirmait ou contestait les données reportées. Nous avons également utilisé des codes pour une demande de clarification ou une relecture de la question et un code général « inadéquate », signifiant que la réponse du répondant ne rentrait dans aucune des catégories de réponse offertes. Les codes de résultat étaient simplement « adéquate » et « inadéquate ».

A. *Codes d'intervieweur*
- S : Standard ; a lu la question telle que rédigée ou avec un changement mineur qui n'a pas modifié la signification.
- MC1 : Énoncé reporté lu comme une question (par exemple : « La dernière fois, vous nous avez dit que vous faisiez de l'hypertension. Est-ce correct ? »)
- MC2 : Question reportée lue comme un énoncé (par exemple : « Et votre date de naissance était le 25 mai 1933. »)
- MC3 : Tout changement qui a modifié ou pourrait modifier la signification de la question.
- O : Omission
- I/O : Enregistrement inaudible ou comportement ne correspondant pas à l'un des codes susmentionnés.

B. *Codes de répondant*
- AR : Adéquate ; a reconnu ou n'a pas contesté les données reportées.
- AC : Adéquate ; a contesté ou mis en doute les données reportées.
- CL : Demande de clarification
- RI : Réponse ou élaboration inadéquate
- NSP : Ne sait pas
- R : Refus
- I/O : Enregistrement inaudible ou comportement ne correspondant pas à l'un des codes susmentionnés.

C. *Codes de résultat*
- AR : Adéquate ; la réponse finale correspond à l'une des catégories de réponse données.
- IA : Inadéquate ; la réponse finale ne correspond à aucune des catégories de réponse.
- NSP : Ne sait pas
- R : Refus
- I/O : Enregistrement inaudible ou comportement ne correspondant pas à l'un des codes susmentionnés.

Figure 2 Codes de comportement

3. Résultats

Nous présenterons d'abord les résultats pour chaque sujet, puis nous discuterons des thèmes pour l'ensemble des sujets. En ce qui concerne le code de résultat, des réponses adéquates ont été obtenues dans la grande majorité des cas (dans plus de 90 % d'entre eux) et les variations d'un item à l'autre étaient faibles, si bien que nous ne présenterons pas ces résultats.

3.1 Items démographiques

Dans le module des caractéristiques démographiques, si nous examinons d'abord le comportement de l'intervieweur, la mesure dans laquelle les intervieweurs respectent la technique standardisée variait assez fortement, allant de 40 % à 79 %, selon l'item (voir le tableau 1A). Il convient de souligner que, dans ce tableau et les autres, la base varie

d'un item à l'autre, pour deux raisons. Premièrement, comme nous l'avons mentionné à la section 2.2, les items RÉSIDE et DDN ont été administrés à l'échantillon de la vague précédente d'ELSA ainsi qu'à l'échantillon de remise à niveau (base de 104 interviews codées), tandis que pour les autres items, seul l'échantillon de la vague précédente d'ELSA a été utilisé (une base de 87 interviews codées). Deuxièmement, l'analyse a été effectuée au niveau de la question (plutôt qu'au niveau de la personne), de sorte que certains items ont été administrés plusieurs fois à une personne particulière. Par exemple, RÉSIDE a été administré pour chaque personne supplémentaire vivant dans le ménage, ENFANT a été administré uniquement si un enfant vivait dans le ménage, et il a été répété pour chaque enfant supplémentaire dans le ménage. En outre, les items sur la santé ont été répétés pour chaque maladie dans une catégorie donnée. Par exemple, si un répondant avait déclaré six types différents de maladie cardiovasculaire lors de la vague précédente, les questions sur la santé ont été répétées six fois pour cette personne, une fois pour chaque maladie dans la catégorie générale des maladies cardiovasculaires.

Comme nous l'avions prédit après notre écoute antérieure (non systématique) des enregistrements, dans la plupart des cas, quand les intervieweurs se sont écartés du texte, ils ont transformé une question en énoncé (par exemple : « Est-ce que la date de naissance de NOM est le 1er janvier ? » devenait « Et la date de naissance de NOM est le 1er janvier. ») Ce comportement s'est produit dans 33 % à 37 % des cas pour les deux premiers items (RÉSIDE et DDN), et seulement dans 8 % des cas pour le dernier item (ENFANT), ce qui n'est peut-être pas si étonnant étant donné la nature des items. Les réponses aux deux premiers d'entre eux peuvent sembler évidentes − particulièrement à la troisième vague − et les intervieweurs ont peut-être hésité à poser une question dont la réponse était évidente. En effet, l'item RÉSIDE a été tout bonnement omis 18 % du temps, peut-être parce que l'intervieweur s'adressait à la personne à laquelle il était fait référence dans la question. Par contre, le troisième item visait à recueillir des renseignements au sujet d'une autre personne dans le ménage (un enfant), l'information était assez précise (nom et date de naissance) et la question réelle (« Ces détails sont-ils corrects ? ») ne paraissait peut-être pas avoir une réponse évidente. Autrement dit, elle pourrait paraître plus « légitime » que demander à une personne, dans un endroit qui semble être son domicile, « Vivez-vous encore ici ? » Cela pourrait expliquer pourquoi le dernier item a été lu tel que rédigé si fréquemment − dans 79 % des cas.

Pour ce qui est du comportement des répondants, dans l'ensemble, ces derniers ont fourni directement une réponse codable dans plus de 80 % des cas (voir le tableau 1B). (Notons que, dans certains cas, la base pour le comportement des répondants pour n'importe quel item est plus petite que celle pour le comportement de l'intervieweur pour le même item. Cette situation est due à une combinaison d'intervieweurs omettant l'item (auquel cas il n'y avait pas de comportement du répondant à coder) et de données manquantes. Les répondants ont rarement contesté les données reportées (proportion allant jusqu'à 5 % des cas seulement), et la plupart des contestations étaient dues à des erreurs de saisie-clavier dans la date de naissance ou le nom enregistré antérieurement.

Tableau 1A
Comportement de l'intervieweur pour les items démographiques

Item	Base (n)	Code de comportement de l'intervieweur (en %)				
		Question lue telle que rédigée	Question lue comme un énoncé	Autre changement majeur	Omise	Autre
RÉSIDE : NOM réside-t-il/elle encore ici ?	120	40	33	4	18	5
DDN : Puis je simplement vérifier, est ce que la date de naissance de NOM est [DDN] ?	107	57	37	1	1	4
ENFANT : Nos dossiers indiquent qu'au moment de la dernière interview, vous aviez un enfant nommé NOM, dont la date de naissance est [DDN]. Ces détails sont-ils corrects ?	84	79	8	11	0	2

3.2 Items sur la santé

Comme nous l'avons mentionné plus haut, les questions sur la santé ont été posées en deux parties. D'abord, on a lu un énoncé au sujet de l'affection déclarée au cours de la vague précédente, puis on a posé une question pour déterminer si elle existait encore. Le niveau global de « lecture exacte » de ces items était moyen − variant de 41 % à 76 %, mais se situait généralement à la limite inférieure des 60 % (voir le tableau 2A). Quand les intervieweurs s'écartaient du texte, ils avaient tendance à transformer l'énoncé en une question (20 % à 38 % du temps) en ajoutant quelque chose du genre « Cela est-il correct ? » à la fin de l'énoncé. Les intervieweurs omettaient alors souvent tout bonnement de poser la question réelle « Avez-vous encore cette affection ? » − dans 13 % à 18 % des cas. Les conséquences sont importantes ici, parce que cela signifie que le répondant reçoit une question fondamentalement différente, plus précisément « Est-il correct que vous avez déclaré cette affection la dernière fois ? » au lieu de « Avez-vous encore cette affection maintenant ? »

Un autre problème s'est manifesté quand la question réelle « Avez-vous encore [l'affection X] ? » était lue, les intervieweurs lisant souvent un énoncé plutôt qu'une question − « Et vous l'avez encore. » − dans 3 % à 16 % des cas. Cette situation a également des conséquences importantes en ce qui concerne la qualité des données, parce qu'il n'est pas donné au répondant l'occasion de réfléchir s'il a vraiment encore l'affection ; il lui est simplement dit qu'il l'a encore.

En ce qui concerne le comportement des répondants (tableau 2B), le niveau de comportement adéquat était assez élevé − plus de 90 % pour les maladies cardiovasculaires et chroniques, et 72 % pour les maladies des yeux (cependant, ici, la base n'était que de 21 cas). Les répondants ont contesté les données de la vague précédente pour diverses raisons. Certains ont dit qu'ils avaient souffert de la maladie, mais qu'ils ne l'avaient plus, et c'est essentiellement la façon dont le questionnaire était censé fonctionner. Toutefois, dans d'autres cas, les données reportées étaient problématiques ; les répondants ont nié avoir déclaré la maladie à la vague précédente ou ils n'étaient pas d'accord avec la description de la maladie. Par exemple, dans un cas, une maladie avait été enregistrée comme un cancer à la vague précédente et quand on lui a posé des questions à ce sujet lors de la vague suivante, le répondant a déclaré qu'il ne s'agissait pas d'un cancer. Il n'était pas certain du diagnostic, mais a déclaré que ce n'était pas un cancer. Dans un autre cas, un répondant avait déclaré avoir un trouble de mémoire à la vague précédente, mais cet état particulier était regroupé avec d'autres maladies connexes dans le questionnaire (« démence, sénilité ou trouble de mémoire »). Quand la question d'IRI est apparue à l'écran, l'intervieweur a lu seulement « démence » et le répondant a réfuté l'information. Ce n'est que quand l'intervieweur est retourné en arrière et qu'il a lu la question en entier, en énumérant les trois états, que le répondant a confirmé qu'il avait un trouble de mémoire. Enfin, dans un cas, la présence d'autres membres de la famille semblait poser un problème. Par exemple, quand il a été indiqué à un répondant qu'il avait déclaré une certaine maladie chronique à la vague précédente, il a répondu « Vraiment ? » et sa femme a répondu « oui ».

3.3 Item sur le véhicule

L'item sur le véhicule était semblable aux items sur la santé, en ce sens qu'il comportait d'abord un énoncé au sujet de ce qui avait été enregistré à la vague précédente, puis une question pour savoir si la situation était encore la même. Toutefois, une différence importante était qu'au lieu de présenter l'énoncé et la question sous forme de deux items distincts sur deux écrans différents, ils étaient regroupés en un seul item. De tous les items du questionnaire, celui sur le véhicule est celui pour lequel le niveau de lecture de la question par l'intervieweur telle qu'elle était rédigée a été le plus élevé, soit 82 % (voir le tableau 3A). Les problèmes relevés dans le module sur la santé − les intervieweurs transformant l'énoncé en une question, ou la question en un énoncé, ou omettant la question − ne se sont pas manifestés très souvent ici, peut-être parce que le style d'IRI était différent. Plus précisément, les intervieweurs n'avaient pas à lire un énoncé au sujet de la réponse antérieure, mais pouvaient passer directement à la question : « Possédez-vous encore ce véhicule ? » En n'affichant pas l'énoncé sur les données de la vague antérieure comme un item distinct, les intervieweurs ont peut-être été moins tentés de transformer celui-ci en une question en demandant, par exemple, « Est-il correct que vous avez déclaré ce véhicule la dernière fois ? » Par conséquent, la question prévue − celle de savoir si la personne possédait encore le véhicule − a été posée, plutôt qu'une question non prévue (« Avez-vous déclaré que vous possédiez ce véhicule la dernière fois ? »). Cependant, parmi les comportements non standard, il y a encore eu plusieurs cas (8 % du temps) où les intervieweurs ont transformé la question en un énoncé : « Et vous possédez encore le véhicule xx. » Cela pourrait tenir au fait que l'intervieweur avait vu le véhicule en question en approchant du domicile du répondant.

Le comportement des répondants était semblable ici à celui observé pour le module sur la santé. Ils ont fourni directement une réponse codable dans 80 % des cas (voir le tableau 3B). Ils ont rarement contesté les données reportées (6 % des cas) et les raisons étaient variées. L'un était dû à des erreurs de saisie-clavier de l'information reportée sur l'immatriculation et deux étaient basés sur un changement réel (dans un cas, le répondant possédait une automobile différente et dans l'autre, il avait cessé de conduire).

Tableau 1B
Comportement des répondants pour les items démographiques

Item	Base (n)	Code de comportement du répondant (en %)				
		Adéquate ; a confirmé les données reportées	Adéquate ; a contesté les données reportées	Clarification	Inadéquate	Autre
RÉSIDE : NOM réside-t-il/elle encore ici ?	91	81	1	1	4	12
DDN : Puis je simplement vérifier, est ce que la date de naissance de NOM est [insérer la date de naissance (DDN)] ?	102	91	0	0	0	9
ENFANT : Nos dossiers indiquent qu'au moment de la dernière interview, vous aviez un enfant nommé NOM, dont la date de naissance est [DDN]. Ces détails sont ils corrects ?	84	89	5	1	1	4

Tableau 2A
Comportement des intervieweurs pour les items sur la santé

Item	Base (n)	Code de comportement de l'intervieweur (en %)					
		Question lue telle que rédigée	Question lue comme un énoncé	Énoncé lu comme une question	Autre changement majeur	Omise	Autre
PRÉCÉDENTE-YEUX : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	21	62	na	38	0	0	0
ENCORE-YEUX : Avez-vous encore [maladie] ?	19	63	16	na	5	16	0
PRÉCÉDENTE-MCV : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	100	63	na	20	17	0	0
ENCORE-MCV : Avez-vous encore [maladie] ?	79	76	3	na	8	13	1
PRÉCÉDENTE-CHRON : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	59	41	na	34	17	5	3
ENCORE-CHRON : Avez-vous encore [maladie] ?	51	61	14	na	2	18	6

Tableau 2B
Comportement des répondants pour les items sur la santé

Item	Base (n)	Code de comportement du répondant (en %)					
		Adéquate ; a confirmé les données reportées	Adéquate ; a contesté les données reportées	Adéquate*	Inadéquate	Clarification	Autre
PRÉCÉDENTE-YEUX : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	21	62	10	[72]	10	5	15
ENCORE-YEUX : Avez-vous encore [maladie] ?	16	na	na	94	0	0	6
PRÉCÉDENTE-MCV : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	100	87	5	[93]	4	0	4
ENCORE-MCV : Avez-vous encore [maladie] ?	67	na	na	69	24	0	8
PRÉCÉDENTE-CHRON : Nos dossiers indiquent qu'à l'occasion de la dernière interview en janvier 2004, vous avez dit que vous aviez (ou qu'un médecin vous avait dit que vous aviez) une [maladie].	53	85	4	[89]	2	0	9
ENCORE-CHRON : Avez-vous encore [maladie] ?	35	na	na	89	6	0	6

* Pour les items « PRÉCÉDENTE-XX », cette colonne donne la somme de « Adéquate ; a confirmé les données reportées » et « Adéquate ; a contesté les données reportées »

Tableau 3A
Interviewer behavior for vehicle item

Item	Base (n)	Code de comportement de l'intervieweur (en %)				
		Question lue telle que rédigée	Question lue comme un énoncé	Autre changement majeur	Omise	Autre
VÉHICULE : La dernière fois que nous vous avons vu(e), vous nous avez dit que vous étiez le principal utilisateur d'un(e) [MARQUE DU VÉHICULE], avec une immatriculation [LETTRE]. Possédez-vous encore ce véhicule ?	51	82	8	2	4	4

Tableau 3B
Comportement des répondants pour l'item sur le véhicule

Item	Base (n)	Code de comportement du répondant (en %)				
		Adéquate ; a confirmé les données reportées	Adéquate ; a contesté les données reportées	Clarification	Inadéquate	Autre
VÉHICULE : La dernière fois que nous vous avons vu(e), vous nous avez dit que vous étiez le principal utilisateur d'un(e) [MARQUE DU VÉHICULE], avec une immatriculation [LETTRE]. Possédez-vous encore ce véhicule ?	49	74	6	10	4	6

3.4 Interaction entre l'intervieweur et le répondant

En plus de l'analyse par item, nous avons examiné la relation entre le comportement de l'intervieweur et celui du répondant pour l'ensemble des items. Que l'intervieweur lise la question telle que rédigée, qu'il lise les questions comme des énoncés ou qu'il lise les énoncés comme des questions, nous avons constaté que le répondant fournissait une réponse adéquate (et affirmative) dans 87,5 % des cas. Pour d'autres comportements, la taille de la cellule était trop faible pour effectuer une analyse significative.

4. Sommaire et recommandations

La mesure dans laquelle les intervieweurs respectaient les textes normalisés variait assez bien − les questions étaient lues telles que rédigées dans 40 % à 82 % des cas, selon l'item. Quand les intervieweurs s'écartaient du texte, la façon dont ils le modifiaient variait selon le sujet et le style d'IRI, lequel, malheureusement, était confondu, parce que chaque item possédait un style unique. Dans le cas des items démographiques et sur le véhicule, dans la plupart des cas, les intervieweurs transformaient la question en un énoncé (« NOM réside-t-il/elle encore ici ? » devenait « Et NOM réside encore ici. »). Dans le module sur la santé, les intervieweurs lisaient les énoncés sur ce qui avait été déclaré la dernière fois comme des questions. Au lieu de lire simplement l'énoncé « La dernière fois, vous avez déclaré la maladie X », ils ajoutaient « Est-ce correct ? » (qui est une question ambiguë) et omettaient souvent la question « Avez-vous encore la maladie X ? » Par conséquent, la question voulue − c'est-à-dire déterminer si la maladie existait encore − était obscurcie ou omise.

Dans la plupart des cas, les répondants fournissaient des réponses codables au moment du premier échange − dans 72 % à 94 % des cas. Les répondants contestaient assez peu souvent les données reportées (dans 0 % à 10 % des cas), mais quand ils le faisaient, les raisons étaient variées. Certains confirmaient la déclaration de la vague précédente, mais disaient qu'ils ne souffraient plus de la maladie. Certains niaient avoir fait la déclaration au moment de la vague précédente, et certains n'étaient pas d'accord avec les détails des données reportées. Il convient de souligner que ce premier scénario est celui que l'instrument était destiné à susciter, si bien que dire que le répondant a « contesté » la déclaration antérieure est en fait une erreur d'appellation. Ici, les répondants ne contestent pas ce qu'ils ont dit antérieurement, mais confirment plutôt leur déclaration antérieure, puis font part d'un changement. Cependant, au moment où nous avons établi le cadre de codage, nous n'avions entendu les répondants contester les données antérieures que dans très peu de cas ; la majorité d'entre eux étaient simplement d'accord avec les données reportées. Par conséquent, nous ne nous sommes pas rendu compte qu'il aurait été utile de créer des codes distincts pour le fait d'être d'accord avec les données reportées et pour la déclaration de changements réels par opposition à la contestation effective de la déclaration antérieure. Cependant, même dans l'ensemble de données complet, la fréquence à laquelle les répondants n'étaient pas tout simplement d'accord avec les données de la vague antérieure était trop faible pour une riche analyse et un plus grand ensemble de données sera nécessaire pour aborder cette question.

Pour ce qui est des recommandations, les résultats susmentionnés suggèrent fortement que les concepteurs de questionnaires devraient éviter de formuler des énoncés des données de la vague précédente sans question réelle, parce que les intervieweurs sont trop tentés de transformer ces énoncés en question, ce qui obscurcit la question de savoir si la situation de la vague antérieure existe encore. S'il est important de confirmer ou de vérifier l'information fournie à l'occasion d'une vague antérieure, il faut le faire de manière explicite en ajoutant une question discrète à l'énoncé, tel que : « La dernière fois, j'ai enregistré que vous aviez la maladie X. Est-ce correct ? » Des questions subséquentes pourraient alors être posées pour déterminer si la maladie existe encore. Séparer les deux concepts de cette façon communiquerait au répondant qu'il s'agit de deux questions distinctes, l'une destinée à savoir si la déclaration antérieure a été enregistrée correctement et l'autre, si la maladie existe encore. Si les chercheurs n'ont pas de raison logique de vouloir confirmer l'exactitude des données enregistrées précédemment, une approche plus efficace consisterait à demander : « La dernière fois, j'ai inscrit que vous aviez la maladie X. Avez-vous encore la maladie X ? »

Nos résultats au sujet du module sur la santé donnent à penser que, pour certains sujets, il est important de donner les données de la vague précédente en retour autant que possible dans les propres mots du répondant. Quand la description que ce dernier avait faite de sa maladie était obscurcie parce que le questionnaire ou l'intervieweur regroupait cette maladie avec d'autres affections, le répondant ne reconnaissait plus la maladie qu'il avait déclarée originalement.

Enfin, nos résultats suggèrent une recommandation plus générale, à savoir que le style d'IRI devrait être adapté soigneusement en fonction de l'item en question. Par exemple, pour des sujets pour lesquels la réponse est peu susceptible de changer d'une vague à l'autre (comme la date de naissance), il faut éviter de reposer les questions, parce que les intervieweurs les lisent souvent comme des énoncés ou les omettent tout bonnement. Pour ces sujets, il serait peut-être plus efficace de vérifier explicitement l'exactitude de la déclaration antérieure (comme il est proposé plus haut)

ou d'éviter de reporter l'information. Une approche hybride pour une étude comportant plusieurs vagues consisterait à vérifier l'exactitude des données enregistrées antérieurement à la deuxième vague, puis d'accepter les données comme étant correctes et d'éviter tout simplement de les reconfirmer au moment de vagues ultérieures.

Le présent rapport est diffusé afin d'informer les parties intéressées des travaux de recherche en cours et de favoriser la discussion. Les opinions exprimées concernant les questions méthodologiques sont celles des auteurs et ne reflètent pas nécessairement celles du US Census Bureau ou du National Centre for Social Research.

Remerciements

Les auteurs remercient Carli Lessof du soutien constant qu'elle a offert pour leurs travaux de recherche sur les méthodes d'enquête et Jeff Moore, des révisions des ébauches antérieures du présent article. Nous remercions aussi Hayley Cheshire, Michelle Gray, Fiona Andrews, Lizzie Hacker, Colin Miceli, Marian Bolden, Audrey Hale et Laura Street de leur aide pour la collecte et le traitement des données. Cette étude a été financée par le National Centre for Social Research et l'étude ELSA est financée par le National Institute of Ageing (aux États-Unis) et par un consortium de ministères du Royaume-Uni coordonné par l'Office for National Statistics.

Bibliographie

Jäckle, A., et Lynn, P. (2004). Dependent Interviewing and Seam Effects in Work History Data. Document de travail #2005-24 de l'Institute for Social and Economic Research, Colchester, Royaume-Uni : University of Essex.

Moore, J.C., Bates, N., Pascale, J. et Okon, A. (2006). Tackling Seam Bias Through Questionnaire Design. Communication sollicitée présentée à la Conférence sur la méthodologie des enquêtes longitudinales, 12 au 14 juillet, Essex, Royaume-Uni.

Pascale, J., et Mayer, T. (2004). Alternative methods for exploring confidentiality issues related to dependent interviewing. *Journal of Official Statistics*, 20, 2, 357-377.

Polivka, A., et Rothgeb, J. (1993). Redesigning the CPS Questionnaire. Monthly Labor Review, 10 au 28 septembre.

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 169-179
Statistique Canada, N° 12-001-X au catalogue

Imputation pour les non-réponses non monotones dépendant de la dernière valeur dans les enquêtes longitudinales

Jing Xu, Jun Shao, Mari Palta et Lin Wang [1]

Résumé

Dans les enquêtes longitudinales, la non-réponse survient souvent selon un schéma non monotone. Nous considérons l'estimation des moyennes dépendantes du temps sous l'hypothèse que le mécanisme de non-réponse dépend de la dernière valeur. Puisque cette dernière valeur peut elle-même manquer quand la non-réponse est non monotone, le mécanisme de non-réponse examiné est non ignorable. Nous proposons une méthode d'imputation qui consiste à établir d'abord certains modèles d'imputation par la régression en fonction du mécanisme de non-réponse, puis à appliquer l'imputation par la régression non paramétrique. Nous supposons que les données longitudinales suivent une chaîne de Markov admettant des moments finis de deuxième ordre. Aucune autre contrainte n'est imposée à la distribution conjointe des données longitudinales et à leurs indicateurs de non-réponse. La variance est estimée par une méthode du bootstrap. Nous présentons certains résultats de simulation et un exemple concernant une enquête sur l'emploi.

Mots clés : Bootstrap ; données manquantes non monotones ; dépendante de la dernière valeur ; non-réponse non ignorable ; régression non paramétrique.

1. Introduction

Une enquête est dite longitudinale si les données sont recueillies auprès de chaque unité échantillonnée à plusieurs points dans le temps. Par exemple, dans l'enquête *Current Employment Survey* (CES), habituellement appelée enquête sur la paye, menée par le Bureau of Labor Statistics des États-Unis, les données sont recueillies mensuellement auprès des établissements, par la poste, par téléphone, par télécopieur et par entrée de données électroniques (Butani, Harter et Wolter 1997). L'enquête *Survey of Income and Program Participation* (SIPP) et de nombreuses enquêtes économiques réalisées par le Census Bureau des États-Unis sont d'autres exemples. Des cas de non-réponse surviennent dans les études longitudinales. Nous supposons que chaque unité échantillonnée répond au moment de l'enquête de base (premier point dans le temps). La non-réponse est monotone si une unité non répondante à un certain point dans le temps ne recommence pas à participer à l'enquête par après. Cependant, la non-réponse non monotone est fréquente dans des enquêtes telles que la CES et la SIPP et comporte une plus grande gamme de profils de non-réponse.

Soit $y_1, \ldots, y_T$ les valeurs d'une variable fournies par une unité échantillonnée, où T est le nombre total de points dans le temps, et $\delta_1, \ldots, \delta_T$ les indicateurs de réponse ($\delta_t = 1$ si y_t est une réponse et $\delta_t = 0$ si y_t est une non-réponse). La non-réponse est entièrement aléatoire si les indicateurs $(\delta_1, \ldots, \delta_T)$ sont statistiquement indépendants de $(y_1, \ldots, y_T)$, ce qui est rare dans les enquêtes. Une hypothèse plus raisonnable est que la non-réponse au temps t dépend des valeurs antérieures observées ou non observées $y_1, \ldots, y_{t-1}$. Dans le présent article, nous nous concentrons sur l'hypothèse plus forte que le mécanisme de non-réponse dépend de la dernière valeur, c'est-à-dire que la non-réponse y_t dépend de la dernière valeur y_{t-1} (observée ou non observée). L'hypothèse que le mécanisme de non-réponse dépend de la dernière valeur est formulée dans de nombreuses enquêtes économiques (par exemple la CES ; voir Butani, Harter et Wolter 1997). Si la non-réponse est également monotone, y_t est une non-réponse avec certitude ou y_{t-1} est observée. Il s'agit d'un cas particulier de ce que l'on nomme données manquantes ignorables (Little et Rubin 1987). Toutefois, si la non-réponse est non monotone, la non-réponse dépendante de la dernière valeur est non ignorable, car le fait que y_t soit une réponse dépend du fait que y_{t-1} pourrait être une non-réponse.

Les méthodes existantes de traitement de la non-réponse non monotone se résument brièvement comme il suit. Sous modélisation paramétrique, des méthodes telles que le maximum de vraisemblance, l'imputation multiple ou l'analyse bayésienne peuvent être appliquées (par exemple Troxel, Harrington et Lipsitz 1998 ; Troxel, Lipsitz et Harrington 1998 ; Schafer 1997) s'il est possible de trouver un modèle paramétrique approprié pour la distribution conjointe de $(y_1, \ldots, y_T)$ et $(\delta_1, \ldots, \delta_T)$. La validité de ces méthodes dépend toutefois de la spécification correcte des modèles paramétriques. Une méthode d'imputation simple par la régression linéaire (voir, par exemple, Butani et coll. 1997) consiste à imputer à une non-réponse y_t la valeur prédite par un modèle ajusté de régression linéaire de y_t en fonction de y_{t-1}, où le modèle de régression est ajusté en utilisant des données provenant d'unités échantillonnées

1. Jing Xu, Jun Shao et Lin Wang, Department of Statistics, University of Wisconsin, Madison, WI 53706 ; Mari Palta, Department of Population Health Sciences et Department of Biostatistics and Medical Informatives, University of Wisconsin, Madison, WI 53706.

pour lesquelles y_t ainsi que y_{t-1} sont observées et la prédiction est faite en utilisant pour prédicteur soit la valeur observée y_{t-1} soit la valeur imputée antérieurement d'une non-réponse y_{t-1}. Néanmoins, sous le mécanisme de non-réponse non monotone (1), nous pouvons montrer que l'imputation par la régression linéaire simple produit un biais même si le modèle de régression linéaire de y_t en fonction de y_{t-1} est correct. Le biais est causé principalement par la façon erronée d'utiliser les valeurs imputées y_{t-1} pour imputer les valeurs manquantes y_t. Une méthode de censure permet de créer un ensemble de données où la non-réponse est monotone en écartant toutes les réponses d'une unité échantillonnée après la première valeur de y manquante. Des méthodes appropriées pour la non-réponse monotone (Paik 1997 ; Robins, Rotnitzky et Zhao 1995 ; Troxel, Lipsitz et Brennam 1997) peuvent alors être appliquées à l'ensemble de données réduit. Bien que cette approche produise des estimateurs corrects, elle n'est pas efficace si T n'est pas petit, puisque de nombreuses réponses sont écartées.

L'objectif du présent article est de proposer une méthode d'imputation pour les enquêtes longitudinales pour lesquelles la non-réponse est non monotone et le mécanisme de non-réponse dépend de la dernière valeur (1). L'imputation est une méthode utilisée couramment pour compenser la non-réponse dans les analyses de données d'enquête (Kalton et Kasprzyk 1986 ; Rubin 1987). Après avoir imputé toutes les non-réponses, les paramètres (tels que la moyenne de y_t) peuvent être estimés par des méthodes classiques en traitant les valeurs imputées comme des observations. La méthode d'imputation que nous proposons produit des estimateurs approximativement sans biais et convergents des moyennes de $y_1, \ldots, y_T$.

La présentation de la suite de l'article est la suivante. À la section 2, nous décrivons nos hypothèses. À la section 3, nous exposons le processus d'imputation. À la section 4, nous discutons de certaines propriétés des estimateurs résultants des moyennes de population, ainsi que de la méthode du bootstrap proposée pour estimer la variance. À la section 5, nous présentons certains résultats de simulation. À la section 6, nous donnons un exemple fondé sur la CES. Enfin, à la dernière section, nous résumons l'étude.

2. Hypothèses et modèle d'imputation

Soit P une population finie désignée par l'indice $i = 1, \ldots, N$ et soit S un échantillon de taille n tiré de P conformément à un plan d'échantillonnage donné. Selon le plan d'échantillonnage, les poids de sondage $w_i, i \in S$ sont construits de manière que, pour tout ensemble de valeurs $\{z_i : i \in P\}$,

$$E_s\left(\sum_{i \in S} w_i z_i\right) = \sum_{i=1}^{N} z_i,$$

où E_s est l'espérance par rapport à S. Pour chaque unité $i \in P$, $(y_{1,i}, \ldots, y_{T,i})$ est un vecteur d'items d'intérêt obtenus aux temps $t = 1, \ldots, T$. En présence de non-réponse, chaque unité possède aussi le vecteur $(\delta_{1,i}, \ldots, \delta_{T,i})$ d'indicateurs de réponse. Pour simplifier, nous omettrons l'indice *i* dans notre discussion.

Nous adoptons une approche assistée par modèle en supposant que les vecteurs $(y_{1,i}, \ldots, y_{T,i}, \delta_{1,i}, \ldots, \delta_{T,i})$ sont indépendants et identiquement distribués (i.i.d.) provenant d'une superpopulation. L'hypothèse i.i.d. peut être relâchée en divisant P en plusieurs sous-populations (appelées classes d'imputation) de sorte qu'elle soit vérifiée approximativement dans chaque classe d'imputation. Les classes d'imputation sont habituellement construites en utilisant une variable catégorique dont les valeurs sont observées pour toutes les unités échantillonnées ; par exemple, sous échantillonnage stratifié, les strates ou les unions de strates sont fréquemment utilisées comme classes d'imputation. Chaque classe d'imputation devrait contenir un grand nombre d'unités échantillonnées. Souvent, si de nombreuses strates sont de petite taille, les classes d'imputation sont crées par poststratification (Valliant 1993) et/ou regroupement de petites strates.

Quand les classes d'imputation sont construites, l'imputation est effectuée dans chacune d'elles. Donc, pour simplifier, à partir de maintenant, nous supposons qu'il n'existe qu'une seule classe d'imputation.

Sous le mécanisme de non-réponse dépendant de la dernière valeur,

$$P(\delta_t = 1 \mid y_1, \ldots, y_T, \delta_1, \ldots, \delta_{t-1}, \delta_{t+1}, \ldots, \delta_T) = P(\delta_t = 1 \mid y_{t-1}), \quad t = 2, \ldots, T. \quad (1)$$

Nous ne formulons aucune autre hypothèse concernant $P(\delta_t = 1 \mid y_{t-1})$. En l'absence de non-réponse, nous supposons que $(y_1, ..., y_T)$ est une chaîne de Markov, c'est-à-dire que

$$L(y_t \mid y_1, \ldots, y_T) = L(y_t \mid y_{t-1}), \quad t = 2, \ldots, T, \quad (2)$$

où $L(\xi \mid \zeta)$ désigne la loi conditionnelle de ξ sachant ζ. Nous n'émettons aucune autre hypothèse concernant $L(y_t \mid y_{t-1})$, excepté que y_t possède un moment fini de deuxième ordre. Dans de nombreuses enquêtes économiques, une hypothèse plus forte que (2) est faite, à savoir :

$$y_t = \beta y_{t-1} + \sqrt{|\, y_{t-1} \,|}\, \varepsilon_t, \quad t = 2, \ldots, T, \quad (3)$$

où β est un paramètre inconnu, les ε_t sont indépendants des y_t, $\varepsilon_1 = 0$ et les $\varepsilon_2, \ldots, \varepsilon_T$ ont une moyenne nulle et une variance commune (par exemple les données de la CES ; voir Butani et coll. 1997). Sous (3), le meilleur

estimateur linéaire sans biais de β est l'estimateur par le ratio bien connu.

Pour considérer les résultats asymptotiques, nous adoptons le cadre de travail de Krewski et Rao (1981) et de Bickel et Freedman (1984). Nous supposons que la population finie P est un membre d'une suite de populations finies désignée par l'indice v. On sous-entend que $v \to \infty$ pour tous les processus limites. Quand $v \to \infty$, la taille de la population N et la taille de l'échantillon n augmentent à l'infini. Dans les enquêtes par sondage, on impose habituellement aux poids w_i les conditions de régularité suivantes :

$$n \max_{i \in P} w_i \le b_0 N \qquad \text{et} \qquad n \mathrm{Var}_s \left(\sum_{i \in S} w_i \right) \le b_1 N^2, \qquad (4)$$

où b_0 et b_1 sont des constantes positives et Var_s est la variance par rapport au plan d'échantillonnage. Dans (4), la première condition assure qu'aucun poids w_i ne soit anormalement grand (voir Krewski et Rao 1981), tandis que la deuxième condition signifie que $\mathrm{Var}_s(\sum_{i \in S} w_i / N)$ est, au plus, d'ordre n^{-1}. Ces conditions sont satisfaites pour les plans d'échantillonnage aléatoires simples stratifiés.

3. Processus d'imputation

L'imputation que nous proposons est un type d'imputation par la régression. Donc, l'un des problèmes clés de notre méthode est de trouver les bons « prédicteurs » pour les non-réponses. Pour une non-réponse $y_t, t \ge 2,$ soit r le point dans le temps auquel l'unité a produit la dernière valeur observée, c'est-à-dire que y_r est observée, mais que $y_{r+1}, ..., y_{t-1}, y_t$ sont des non-réponses. Sous les hypothèses (1) et (2), nous pouvons utiliser y_r comme prédicteur pour imputer y_t.

3.1 Cas où $r = t - 1$

Examinons d'abord le cas où $r = t - 1$. Soit

$$\phi_{t,t-1}(y_{t-1}) = E(y_t | y_{t-1}, \delta_t = 0, \delta_{t-1} = 1)$$

l'espérance conditionnelle (fonction de régression) pour une non-réponse y_t avec valeur observée y_{t-1}. Si $\phi_{t,t-1}$ est connue, nous pouvons simplement imputer y_t par $\phi_{t,t-1}(y_{t-1})$. Mais $\phi_{t,t-1}$ est habituellement inconnue. Nous montrons en annexe que l'hypothèse (1) implique que

$$\phi_{t,t-1}(y_{t-1}) = E(y_t | y_{t-1}, \delta_t = 1, \delta_{t-1} = 1), \quad t = 2, \ldots, T. \qquad (5)$$

Donc, $\phi_{t,t-1}$ peut être estimée par régression de y_t sur y_{t-1} en utilisant les données provenant de toutes les unités échantillonnées pour lesquelles y_t et y_{t-1} sont observées.

L'idée d'utiliser (5) pour l'imputation est la même que celle appliquée dans le cas de la non-réponse monotone traité par Paik (1997). Cependant, contrairement au cas de la non-réponse monotone, $\phi_{t,t-1}(x)$ peut ne pas être linéaire en x pour la non-réponse non monotone. Donc, nous considérons la méthode non paramétrique décrite dans Cheng (1994) pour la régression. L'estimateur à noyau de $\phi_{t,t-1}(x)$ est

$$\hat{\phi}_{t,t-1}(x) = \sum_{i \in S} \kappa\left(\frac{x - y_{t-1,i}}{h} \right) w_i I_{t,t-1,i} y_{t,i} \Big/ \sum_{i \in S} \kappa\left(\frac{x - y_{t-1,i}}{h} \right) w_i I_{t,t-1,i},$$

où $\kappa(x)$ est une densité de probabilité, $h > 0$ est une fenêtre et

$$I_{t,t-1,i} = \begin{cases} 1 & \delta_{t,i} = 1, \delta_{t-1,i} = 1 \quad t = 2, \ldots, T. \\ 0 & \text{autrement,} \end{cases}$$

Une non-réponse $y_{t,j}$ avec réponse $y_{t-1,j}$ est imputée par

$$\tilde{y}_{t,j} = \hat{\phi}_{t,t-1}(y_{t-1,j}).$$

Cheng (1994) a proposé la fenêtre $h = Cn^{-2/5}$, où C est une constante. En général, C peut différer d'une application à l'autre et devrait être choisie en utilisant des techniques exposées dans la littérature sur l'estimation par le noyau (par exemple, Cheng 1994 et le chapitre 5 de Härdle 1990) et/ou des études empiriques.

3.2 Cas où $r < t - 1$

Quand $r < t - 1,$ la situation est plus compliquée. Soit

$$\phi_{t,r}(y_r) = E(y_t | y_r, \delta_t = \cdots = \delta_{r+1} = 0, \delta_r = 1).$$

Si le mécanisme de non-réponse (1) est non ignorable, la valeur espérée de y_t sachant y_r avec $r < t - 1$ n'est pas la même pour la valeur observée et la valeur manquante de $y_t,$ ce qui exclut l'utilisation de valeurs observées de y_t comme résultat dans l'estimation $\phi_{t,r}$. Xu (2007) démontre explicitement que

$$\phi_{t,r}(y_r) \neq E(y_t | y_r, \delta_t = 1, \delta_{t-1} = a_{t-1}, \ldots, \delta_{r+1} = a_{r+1}, \delta_r = 1) \qquad (6)$$

où $a_j = 0$ ou $1, j = r + 1, \ldots, t - 1$. Par contre, dans le cas de la non-réponse monotone, les deux membres de (6) sont les mêmes (Paik 1997), si bien que le deuxième membre de (6) peut être utilisé comme modèle d'imputation par la régression et les valeurs observées de y_t peuvent être utilisées pour estimer $\phi_{t,r}$.

Nous devons trouver une espérance conditionnelle de y_t (sachant y_r et une certaine situation de réponse) qui est égale à $\phi_{t,r}(y_r)$ et qui nous permet d'exécuter l'imputation. Nous montrons en annexe que

$$\phi_{t,r}(y_r) = E(y_t | y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1), \quad r = 1, \ldots, t - 2, t = 2, \ldots, T. \qquad (7)$$

Pour estimer $\phi_{t,r}$ en ajustant la régression conformément à (7), nous ne possédons pas de valeurs de y_t observées comme résultats dans la régression, parce que $\delta_t = 0$ pour les unités définies par le deuxième membre de l'équation (7). Si nous effectuons l'imputation séquentielle suivant $r = t-1, t-2, \ldots, 1,$ alors les non-réponses y_t pour les unités pour lesquelles $\delta_{r+1} = 1$ ont déjà été imputées. Donc, nous pouvons utiliser ces valeurs de y_t imputées antérieurement comme résultats dans la régression. Bien que, à chaque point fixe t dans le temps, l'imputation soit exécutée séquentiellement suivant $r = t-1, t-2, \ldots, 1,$ l'imputation pour divers points dans le temps peut être effectuée dans n'importe quel ordre, parce que, en tout point t dans le temps, les valeurs imputées sont utilisées comme résultats au temps t uniquement.

Puisque $\phi_{t,r}$ est habituellement non linéaire, nous utilisons la régression à noyau. Pour $t = 2, \ldots, T$ et $r = t-2, t-3, \ldots, 1,$ l'espérance conditionnelle $\phi_{t,r}(x)$ pour tout x est estimée par

$$\hat{\phi}_{t,r}(x) = \sum_{i \in S} \kappa\left(\frac{x - y_{r,i}}{h}\right) w_i I_{t,r,i} \tilde{y}_{t,i} \Big/ \sum_{i \in S} \kappa\left(\frac{x - y_{r,i}}{h}\right) w_i I_{t,r,i}, \quad (8)$$

où $\tilde{y}_{t,i}$ est une valeur imputée antérieurement et

$$I_{t,r,i} = \begin{cases} 1 & \delta_{t,i} = \cdots = \delta_{r+2,i} = 0, \delta_{r+1,i} = \delta_{r,i} = 1, \\ & r = 1, \ldots, t-2, \quad t = 2, \ldots, T. \\ 0 & \text{autrement,} \end{cases}$$

Une non-réponse $y_{t,j}$ avec dernière réponse $y_{r,j}$ est imputée par

$$\tilde{y}_{t,j} = \hat{\phi}_{t,r}(y_{r,j}).$$

L'hypothèse de chaîne de Markov (2) assure que l'utilisation de valeurs imputées antérieurement $\tilde{y}_{t,i}$ comme résultats dans (8) produise un estimateur asymptotiquement valide de $\phi_{t,r}$ (voir le résultat (11) en annexe).

3.3 Une illustration

Afin d'illustrer le processus d'imputation décrit plus haut et la façon dont les schémas de non-réponse sont groupés en cellules d'imputation, nous considérons l'imputation au temps $t = 4$ (tableau 1). Dans le tableau 1, la direction horizontale correspond à quatre points dans le temps et la direction verticale correspond à divers schémas de non-réponse, où chaque schéma est représenté par un vecteur de dimension 4 constitué de 0 et de 1 avec 0 indiquant une non-réponse et 1 indiquant une réponse. Il existe, en tout, $2^{T-1} = 2^3 = 8$ schémas de non-réponse. Conformément au processus d'imputation décrit plus haut, à l'étape 1, nous considérons les non-réponses au temps 4 avec les dernières réponses au temps 3, qui correspondent aux schémas 3 et 4. Conformément au modèle d'imputation (5), nous ajustons une régression en utilisant les données dans les schémas 7 et 8 indiquées par + (utilisées comme prédicteurs) et par × (utilisées comme résultats) dans le bloc du tableau 1 sous le titre Étape 1. Puis, nous obtenons les valeurs imputées (indiquées par ○) à partir de la régression ajustée en utilisant les données indiquées par * comme prédicteurs dans le bloc sous le titre Étape 1. Ensuite, nous nous concentrons sur le bloc du tableau 1 situé sous le titre Étape 2. Les non-réponses au temps $t = 4$ avec les réponses au temps $t = 2$ sont celles du schéma 2. Suivant le modèle d'imputation (7), nous ajustons une régression en utilisant les données du schéma 3 indiquées par + (utilisées comme prédicteurs) et par ⊗ (valeurs imputées antérieurement utilisées comme résultats). Puis, nous obtenons les valeurs imputées (indiquées par ○) à partir de la régression ajustée en utilisant les données indiquées par * comme prédicteurs. Enfin, nous nous concentrons sur le bloc du tableau 1 sous le titre Étape 3. Les non-réponses au temps $t = 4$ avec les dernières réponses au temps $t = 1$ sont celles du schéma 1. Conformément au modèle d'imputation (7), nous ajustons une régression en utilisant les données du schéma 2 indiquées par + (utilisées comme prédicteurs) et par ⊗ (valeurs imputées antérieurement utilisées comme résultats). Ensuite, nous obtenons les valeurs imputées (indiquées par ○) à partir de la régression ajustée en utilisant les données indiquées par * comme prédicteurs.

Tableau 1
Illustration du processus d'imputation au temps $t = 4$

	Étape 1: $r = 3$ Temps				Étape 2: $r = 2$ Temps				Étape 3: $r = 1$ Temps			
Schéma	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
(1,0,0,0)									*			○
(1,1,0,0)						*		○	+			⊗
(1,1,1,0)			*	○		+		⊗				
(1,0,1,0)			*	○								
(1,0,0,1)												
(1,1,0,1)												
(1,0,1,1)			+	×								
(1,1,1,1)			+	×								

+ : données observées utilisées comme prédicteurs dans l'ajustement de la régression
× : données observées utilisées comme réponses dans l'ajustement de la régression
⊗ : données imputées utilisées comme réponses dans l'ajustement de la régression
* : données observées utilisées comme prédicteurs dans l'imputation
○ : valeurs imputées

4. Estimation des moyennes de population en utilisant les données imputées

Soit $\bar{Y}_t$ la moyenne de population finie au temps t. La moyenne d'échantillon basée sur les données observées et imputées s'écrit

$$\hat{\bar{Y}}_t = \sum_{i \in S} w_i \tilde{y}_{t,i}, \qquad (9)$$

où $\tilde{y}_{t,i}$ est égal à la valeur observée si $\delta_{t,i} = 1$ et est une valeur imputée si $\delta_{t,i} = 0$. Établissons maintenant que, en tant qu'estimateur de la moyenne de population au temps t, $\hat{\bar{Y}}_t$ donné par (9) est convergent et asymptotiquement normal sous le cadre asymptotique décrit à la section 2.

Théorème 1. Formulons les hypothèses (1), (2) et (4), et supposons que nous travaillons dans le cadre asymptotique décrit à la section 2. Supposons en outre que les conditions de régularité suivantes sont appliquées :

(C1) $E(y_t^2) < \infty, t = 1, \ldots, T.$

(C2) $0 < P(I_{t,r} = 1) < 1$ et $E[\sigma_{t,r}^2(y_r)/p_{t,r}(y_r)] < \infty$, où $p_{t,t-1}(x) = P(\delta_t = 1 \mid y_{t-1} = x, \delta_{t-1} = 1)$, $p_{t,r}(x) = P(\delta_t = 0 \mid y_r = x, \delta_r = \delta_{r+1} = 1, \delta_{t-1} = \cdots = \mu_X\ \delta_{r+2} = 0)$, $r = 1, \ldots, t-2$, $\sigma_{t,r}^2(x) = \mathrm{Var}(y_t \mid y_r = x, I_{t,r} = 1)$, $I_{t,r}$ est équivalent à $I_{t,r,i}$ avec les $\delta_{t,i}$ remplacés par les δ_t, $r = 1, \ldots, t-1, t = 2, \ldots, T$.

(C3) $\phi_{t,r}(x)$ et $g_{t,r}(x) = p_{t,r}(x) f_r(x)$ possèdent des dérivées secondes bornées telles que

$$E[\{\sigma_{t,r}^2(y_r) + \phi_{t,r}(y_r)\} g_{t,r}''(y_r) \times \{1 - p_{t,r}(y_r)\} / \sqrt{g_{t,r}(y_r)} < \infty$$

et

$$E[\{\phi_{t,r}(y_r) g_{t,r}''(y_r)^2 + \phi_{t,r}'(y_r) g_{t,r}'(y_r)\} \times \{1 - p_{t,r}(y_r)\} / g_{t,r}(y_r)] < \infty,$$

où $f_r(x)$ est la densité de probabilité de $y_r, r = 1, \ldots, t-1, t = 2, \ldots, T$.

(K) La fonction noyau κ est une densité de probabilité bornée et symétrique sur la ligne réelle avec deuxième moment fini.

(B) La fenêtre h satisfait $nh^2/(\log n)^2 \to \infty$ et $nh^4 \to 0$ quand $n \to \infty$.

Alors, pour $t = 1, \ldots, T$,

$$\sqrt{n}(\hat{\bar{Y}}_t - \mu_t) \to_d N(0, \sigma_t^2), \qquad (10)$$

où $\mu_t = E(y_t)$, σ_t^2 est un paramètre inconnu, et $\to_d$ indique la convergence en loi par rapport à la loi conjointe de $(y_{1,i}, \ldots, y_{T,i}, \delta_{2,i}, \ldots, \delta_{T,i})$ et à l'échantillonnage (modèle et plan).

La preuve est donnée en annexe. les conditions (K) et (B) sont exactement les mêmes que celles énoncées dans Cheng (1994) et les conditions (C1) à (C3) sont les mêmes que celles de Cheng (1994) appliquées aux unités définies par les deuxièmes membres des équations (5) à (7).

À cause de la complexité de l'imputation, il est difficile d'obtenir une forme explicite de σ_t^2 dans (9). Nous considérons la méthode du bootstrap. Un bootstrap correct peut être obtenu en appliquant le processus d'imputation dans chacun des échantillons bootstrap, c'est-à-dire en imputant chaque ensemble de données bootstrap exactement de la même façon que l'ensemble de données originales (Shao et Sitter 1996). Plus précisément, nous procédons comme il suit.

1. Dans chaque classe d'imputation, tirer de l'échantillon un échantillon bootstrap par échantillonnage aléatoire simple avec remise, où la taille de l'échantillon bootstrap est égale au nombre d'unités échantillonnées dans la classe d'imputation. Combiner les échantillons bootstrap pour former S^*. L'ensemble de données bootstrap contient toutes les données observées, les poids, et les indicateurs de réponse des unités dans S^*.
2. Appliquer la méthode d'imputation proposée à l'ensemble de données bootstrap. Calculer l'analogue bootstrap $\hat{\bar{Y}}_t^*$.
3. Répéter indépendamment les étapes qui précèdent B fois pour obtenir $\hat{\bar{Y}}_t^{*1}, \ldots, \hat{\bar{Y}}_t^{*B}$. La variance d'échantillon de $\hat{\bar{Y}}_t^{*1}, \ldots, \hat{\bar{Y}}_t^{*B}$ est notre estimateur de variance par le bootstrap pour $\hat{\bar{Y}}_t$.

Notons que la méthode du bootstrap requiert un grand nombre de calculs répétés, ce qui est le prix à payer pour remplacer le calcul théorique des variances asymptotiques. Il est également possible d'utiliser d'autres méthodes du bootstrap valides pour les données d'enquête décrites dans Shao et Tu (1995, chapitre 6).

À la section suivante, nous évaluons la performance de l'estimateur de variance par le bootstrap proposé à l'aide d'une simulation.

5. Simulation

Nous avons exécuté une étude par simulation pour évaluer la performance de la méthode d'imputation proposée en ce qui concerne l'estimation de la moyenne de y_t. Nous avons considéré un échantillon de taille 1 000. Chaque unité échantillonnée possède des données longitudinales de taille $T = 4$. Les valeurs des moyennes de population pour les quatre points dans le temps sont 1,33, 1,94, 2,73 et 3,67, respectivement. Nous avons produit des données longitudinales conformément à deux modèles. Dans le premier, $(y_1, \ldots, y_T)$ suit une loi normale multivariée et suit le modèle AR(1) avec coefficient de corrélation de 0,9 et erreur-type de 1. Dans le deuxième, $(\log y_1, \ldots, \log y_T)$ suit une loi normale multivariée et suit le modèle AR(1) avec coefficient de corrélation de 0,9 et

erreur-type de 1. Au temps $t = 1$, toutes les données sont observées. Pour $t = 2, \ldots, T$, les non-réponse sont générées en utilisant

$$P(\delta_t = 0 \mid y_{t-1}) = \frac{\exp\{1 - 1,2y_{t-1}\}}{1 + \exp\{1 - 1,2y_{t-1}\}}$$

pour le cas des données normales et

$$P(\delta_t = 0 \mid y_{t-1}) = \frac{\exp\{2 - 0,7y_{t-1}\}}{1 + \exp\{2 - 0,7y_{t-1}\}}$$

pour le cas des données log-normales. Ces modèles de non-réponse ont été choisis de manière que les probabilités inconditionnelles de non-réponse soient environ les mêmes dans les deux cas (voir le tableau 2).

Aux fins de comparaison, nous avons inclus cinq estimateurs dans la simulation : la moyenne d'échantillon des données complètes, qui est utilisée comme cas idéal ; la moyenne d'échantillon des réponses en ne tenant pas compte des non-réponses, la moyenne d'échantillon basée sur l'imputation par la régression linéaire simple décrite à la section 1, la moyenne d'échantillon basée sur la censure (décrite à la section 1) et l'imputation par la régression linéaire pour des données manquantes monotones (Paik 1997), ainsi que la moyenne d'échantillon basée sur la méthode d'imputation que nous proposons. Dans la régression non paramétrique décrite aux sections 3.1 et 3.2, nous avons utilisé la densité normale standard comme noyau $\kappa(x)$ et choisi la fenêtre h de manière qu'elle soit environ $4n^{-2/5}$, c'est-à-dire celle utilisée dans la simulation de Cheng (1994).

Les tableaux 3 et 4 donnent (en se fondant sur 1 000 exécutions de la simulation) le biais relatif et la variance des estimateurs de la moyenne, l'estimateur de variance par le bootstrap (basé sur 200 répliques bootstrap), la probabilité de couverture de l'intervalle de confiance (IC) à 95 % approximatif obtenu en utilisant l'estimateur ponctuel $\pm 1,96 \times \sqrt{\text{variance bootstrap}}$, et la longueur de l'IC. Les résultats des tableaux 3 et 4 se résument comme il suit.

1. Biais. La méthode d'imputation proposée produit des estimateurs dont le biais est négligeable dans tous les cas envisagés. La moyenne d'échantillon des réponses uniquement est manifestement biaisée à moins que $t = 1$. Même si, dans certains cas ses valeurs sont faibles, le biais donne lieu à une probabilité de couverture très faible de l'IC, parce que la variance de la moyenne d'échantillon est beaucoup plus faible que le carré du biais. La méthode d'imputation par la régression linéaire simple n'est correcte que quand $t = 2$ et que les données suivent une loi normale. Dans le cas normal, le biais relatif à $t = 3$ est très petit, mais à $t = 4$, il est de 1,1 %, ce qui donne une probabilité de couverture de l'IC de 76,3 % seulement. La méthode de censure et d'imputation par la régression linéaire est correcte dans le cas normal et produit un faible biais. Par contre, dans le cas log-normal, l'imputation par la régression linéaire simple ainsi que la méthode de censure avec imputation par régression linéaire sont biaisées, parce que les fonctions de régression ne sont pas linéaires.

2. Bootstrap et IC. La méthode du bootstrap comme estimateur de la variance de l'estimateur de la moyenne donne des bons résultats dans tous les cas, même quand l'estimateur de la moyenne comporte un biais. L'IC connexe a une probabilité de couverture qui s'approche du niveau nominal de 95 % quand l'estimateur de la moyenne est sans biais.

3. Imputation proposée contre censure. Quand nous utilisons la censure et l'imputation par la régression linéaire, l'estimateur de la moyenne est biaisé dans le cas log-normal et, donc, la méthode d'imputation proposée est clairement meilleure. Dans le cas normal, les deux méthodes sont correctes. Cependant, les résultats du tableau 3 illustrent l'effet du rejet des données observées. Quand $t = 2$, la méthode de censure donne de meilleurs résultats que la méthode d'imputation proposée, parce qu'aucune unité n'est vraiment censurée et que la méthode de censure s'appuie sur la régression linéaire correcte, tandis que la méthode d'imputation proposée repose sur l'ajustement d'une régression non paramétrique. Quand $t = 3$, la performance de la méthode de censure est environ la même que celle de la méthode d'imputation proposée, mais quand $t = 4$, les résultats de la méthode de censure sont beaucoup moins bons que ceux de la méthode d'imputation proposée. L'examen du tableau 2 montre qu'en moyenne, 25 % des unités échantillonnées sont censurées quand $t = 3$ et que 45 % le sont quand $t = 4$. Le gain lié à l'utilisation de la régression linéaire correcte ne suffit pas à compenser l'effet du rejet des données observées, surtout quand $T = 4$.

Tableau 2
Probabilités inconditionnelles de non-réponse dans l'étude par simulation

	Schéma de non-réponse		Probabilité de non-réponse			
			Cas normal		Cas log-normal	
$t = 3$	Monotone	(1,0,0)	0,14		0,16	
		(1,1,0)	0,12	0,26	0,09	0,25
	Intermittente	(1,0,1)	0,25	0,25	0,20	0,20
	Complète	(1,1,1)	0,49	0,49	0,55	0,55
$t = 4$	Monotone	(1,0,0,0)	0,04		0,06	
		(1,1,0,0)	0,02		0,02	
		(1,1,1,0)	0,04	0,10	0,02	0,10
	Intermittente	(1,0,0,1)	0,10		0,10	
		(1,0,1,0)	0,04		0,03	
		(1,0,1,1)	0,21		0,17	
		(1,1,0,1)	0,10	0,45	0,06	0,36
	Complète	(1,1,1,1)	0,45	0,45	0,54	0,54

Tableau 3
Résultats des simulations pour l'estimation des moyennes (cas normal)

Méthode	Quantité	$t = 1$	$t = 2$	$t = 3$	$T = 4$
Données complètes	Biais relatif	0,0 %	0,0 %	0,0 %	0,0 %
	Variance $\times 10^3$	0,962	0,981	1,052	1,033
	Variance par le bootstrap	1,002	1,002	1,002	1,006
	Prob. de couverture de l'IC	95,4 %	94,9 %	94,5 %	94,5 %
	Longueur de l'IC	0,124	0,124	0,124	0,124
Réponses	Biais relatif		16,8 %	8,3 %	3,5 %
	Variance $\times 10^3$		1,319	1,240	1,051
	Variance par le bootstrap		1,364	1,178	1,062
	Prob. de couverture de l'IC		0,0 %	0,0 %	2,4 %
	Longueur de l'IC		0,145	0,134	0,128
Imputation par la régression linéaire simple	Biais relatif		0,0 %	0,0 %	1,1 %
	Variance $\times 10^3$		1,121	1,434	1,185
	Variance par le bootstrap		1,172	1,466	1,192
	Prob. de couverture de l'IC		94,9 %	94,7 %	76,3 %
	Longueur de l'IC		0,134	0,150	0,135
Censure et imputation par la régression linéaire	Biais relatif		0,0 %	0,0 %	0,0 %
	Variance $\times 10^3$		1,121	1,437	1,642
	Variance par le bootstrap		1,172	1,476	1,819
	Prob. de couverture de l'IC		94,9 %	94,7 %	96,1 %
	Longueur de l'IC		0,134	0,150	0,167
Imputation proposée	Biais relatif		0,2 %	0,3 %	0,2 %
	Variance $\times 10^3$		1,196	1,438	1,264
	Variance par le bootstrap		1,231	1,401	1,224
	Prob. de couverture de l'IC		95,0 %	93,7 %	94,1 %
	Longueur de l'IC		0,137	0,146	0,137

Tableau 4
Résultats des simulations pour l'estimation des moyennes (cas log-normal)

Méthode	Quantité	$t = 1$	$t = 2$	$t = 3$	$t = 4$
Données complètes	Biais relatif	0,0 %	0,0 %	0,0 %	0,0 %
	Variance	0,069	0,172	0,383	1,074
	Variance par le bootstrap	0,067	0,161	0,418	1,138
	Prob. de couverture de l'IC	94,4 %	93,8 %	94,9 %	94,6 %
	Longueur de l'IC	1,008	1,557	2,511	4,142
Réponses	Biais relatif		28,1 %	18,8 %	10,8 %
	Variance		0,366	0,614	1,378
	Variance par le bootstrap		0,344	0,668	1,461
	Prob. de couverture de l'IC		0,1 %	2,1 %	31,6 %
	Longueur de l'IC		2,267	3,171	4,690
Imputation par la régression linéaire simple	Biais relatif		7,0 %	12,6 %	12,5 %
	Variance		0,266	0,877	1,589
	Variance par le bootstrap		0,240	0,807	1,611
	Prob. de couverture de l'IC		71,6 %	39,3 %	23,2 %
	Longueur de l'IC		1,894	3,481	4,938
Censure et imputation par la régression linéaire	Biais relatif		7,0 %	12,1 %	13,8 %
	Variance		0,266	0,874	2,735
	Variance par le bootstrap		0,240	0,836	2,617
	Prob. de couverture de l'IC		71,6 %	43,9 %	36,4 %
	Longueur de l'IC		1,894	3,540	6,277
Imputation proposée	Biais relatif		0,1 %	0,1 %	0,1 %
	Variance		0,189	0,447	1,119
	Variance par le bootstrap		0,179	0,482	1,236
	Prob. de couverture de l'IC		94,5 %	95,7 %	95,6 %
	Longueur de l'IC		1,644	2,697	4,317

6. Exemple

Dans le cadre de la CES présentée à la section 1, les données sur l'emploi sont recueillies mensuellement auprès des établissements non agricoles. Durant tout mois après le mois de référence, le taux de non-réponse est de l'ordre de 20 à 40 % et la non-réponse est non monotone. En outre, il est habituellement supposé que les hypothèses (1) et (2) sont vérifiées. En fait, l'hypothèse (3), qui est plus forte que l'hypothèse (2), est souvent émise (Butani, Harter et Wolter 1997). Nous considérons un échantillon aléatoire simple stratifié tiré d'un ensemble de données de la CES (un sous-ensemble d'un échantillon des années 1980). Le tableau 5 donne la taille des strates, la taille d'échantillon selon la strate, ainsi que le taux de non-réponse selon la strate. Chaque méthode d'imputation est appliquée dans un groupe de strates (groupe 1 = strates 1 à 4 ; groupe 2 = strates 5 à 7 ; groupe 3 = strates 8 à 11 ; groupe 4 = strate 12 ; groupe 5 = strates 13 à 15 ; groupe 6 = strate 16).

Afin d'estimer les nombres d'emplois du mois 1 (référence) au mois 8, nous avons appliqué cinq méthodes dans l'étude par simulation de la section 5. Le noyau et la fenêtre utilisés pour la régression non paramétrique étaient les mêmes que dans la simulation (section 5). Comme les nombres d'emplois dans la population sont tirés une fois par an des dossiers administratifs de l'assurance-chômage, les données correspondant aux non-réponses survenues n'importe quel mois deviennent en fait disponibles plus tard, de sorte que la moyenne d'échantillon fondée sur les données complètes peut être calculée comme norme. Les moyennes d'échantillon calculées par diverses méthodes sont présentées au tableau 6, ainsi que les estimations bootstrap de leur variance (fondées sur une taille d'échantillon bootstrap de 200). Dans le cas des moyennes d'échantillon basées sur les réponses et trois méthodes d'imputation, nous avons également calculé le biais relatif estimé défini comme étant égal à (moyenne d'échantillon/ moyenne d'échantillon basée sur les données complètes) −1.

Le tableau 6 montre que la moyenne d'échantillon basée sur la méthode d'imputation proposée est fort comparable à celle fondée sur les données complètes, tandis que la moyenne d'échantillon basée sur les répondants présente manifestement un biais. Vu que la non-réponse est non ignorable, l'imputation par la régression linéaire simple produit un certain biais à partir du mois 4, quoique le biais relatif estimé soit, au plus, égal à 5,5 % en valeur absolue. La méthode de censure avec imputation par la régression linéaire est entachée d'un certain biais après le mois 4, probablement parce que les données ne suivent pas une loi normale, si bien que l'ajustement de la régression linéaire n'est pas correct. En outre, les variances estimées sont grandes comparativement à celles obtenues pour la méthode proposée, ce qui traduit l'effet du rejet des données observées.

Tableau 5
Taille de la strate, taille de l'échantillon et taux de non-réponse dans l'exemple de la CES

Strate	Taille de la strate	Taille de l'échantillon	Non-réponse en pourcentage							
			$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
1	223	102	0	32,4	39,2	34,3	27,5	30,4	28,4	33,3
2	1 649	110	0	28,2	31,8	30,0	34,5	33,6	26,4	29,1
3	1 900	120	0	37,5	39,2	34,2	44,2	41,7	40,0	40,8
4	419	98	0	41,8	48,0	35,7	38,8	44,9	38,8	38,8
5	1 947	132	0	37,1	33,3	25,8	25,0	27,3	23,5	28,8
6	2 391	180	0	41,1	36,1	42,8	37,8	39,4	39,4	38,3
7	5 365	256	0	35,2	34,0	33,6	36,7	35,2	40,6	39,1
8	2 330	201	0	30,3	36,8	40,3	34,8	37,3	37,3	37,8
9	1 164	113	0	35,4	29,2	33,6	30,1	29,2	32,7	33,6
10	593	106	0	37,7	44,3	40,6	47,2	41,5	37,7	32,1
11	2 222	182	0	24,2	26,4	27,5	27,5	28,0	20,3	27,5
12	6 880	512	0	40,0	39,6	40,6	41,0	41,4	39,8	38,9
13	2 373	160	0	36,9	40,6	36,2	33,8	39,4	30,0	36,9
14	50	42	0	40,5	38,1	28,6	45,2	33,3	33,3	31,0
15	4 100	241	0	36,5	38,6	34,4	34,9	42,7	33,2	32,8
16	3 951	412	0	37,9	36,9	36,9	38,1	40,3	40,5	39,3

Tableau 6
Estimations dans l'exemple de la CES

Méthode	Quantité	$t = 1$	$t = 2$	$t = 3$	$T = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$
Données complètes	moyenne	38,05	38,41	38,70	38,95	39,16	38,91	38,79	38,81
	$\widehat{\text{var.}}$	0,805	0,814	0,828	0,830	0,990	0,832	0,822	0,852
Réponses	moyenne		54,03	54,31	54,15	55,08	55,15	54,20	54,50
	$\widehat{\text{biais}}$		40,7 %	40,3 %	39,0 %	40,7 %	41,7 %	39,7 %	40,4 %
	$\widehat{\text{var.}}$		1,647	1,488	1,506	1,990	1,708	1,413	1,491
Imputation par la régression linéaire simple	moyenne		38,45	38,72	39,33	38,81	41,04	40,54	39,79
	$\widehat{\text{biais}}$		0,1 %	0,1 %	1,0 %	-0,9 %	5,5 %	4,5 %	2,5 %
	$\widehat{\text{var.}}$		0,821	0,834	0,866	0,963	1,979	1,465	1,008
Censure et imputation par la régression linéaire	moyenne		38,45	38,71	39,17	38,25	40,46	40,34	40,30
	$\widehat{\text{biais}}$		0,1 %	0,0 %	0,6 %	-2,3 %	4,0 %	4,0 %	3,8 %
	$\widehat{\text{var.}}$		0,821	0,833	0,881	1,289	1,497	1,630	1,660
Imputation proposée	moyenne		38,37	38,68	38,97	39,10	39,05	38,72	38,88
	$\widehat{\text{biais}}$		-0,1 %	-0,1 %	0,0 %	-0,1 %	0,4 %	-0,2 %	0,2 %
	$\widehat{\text{var.}}$		0,813	0,834	0,837	1,019	0,962	0,924	0,910

$\widehat{\text{biais}}$: (moyenne d'échantillon/moyenne d'échantillon basée sur les données complètes) - 1

$\widehat{\text{var.}}$: estimation bootstrap de la variance

7. Conclusion

Dans le cas de données longitudinales avec non-réponse non monotone, nous proposons une méthode d'imputation sous les hypothèses que les mécanismes de non-réponse dépendent de la dernière valeur et que les données longitudinales suivent une chaîne de Markov. Notre méthode est non paramétrique et produit des estimateurs convergents et suivant asymptotiquement une loi normale des moyennes de population. Comme les variances asymptotiques des estimateurs des moyennes de population sont très compliquées, nous proposons une simple méthode du bootstrap pour estimer la variance. Certains résultats de simulation montrent que la méthode proposée donne de bons résultats. À titre d'illustration, nous utilisons la CES, qui est l'un des exemples qui a motivé notre étude.

En général, la non-réponse au temps t peut dépendre non seulement de la dernière valeur observée au temps $t-1$, mais aussi d'autres valeurs antérieures recueillies à des points dans le temps antérieurs à $t-1$. De surcroît, les données longitudinales ne suivent pas nécessairement une chaîne de Markov, si bien qu'il pourrait exister une dépendance de long terme entre les données provenant de divers points dans le temps. Dans l'un ou l'autre cas, la méthode que nous proposons n'est pas applicable. Une méthode générale est en cours d'élaboration.

Annexe

Preuve de (5)

Soit $L(\xi)$ la distribution de ξ et $L(\xi \mid \zeta)$ la distribution conditionnelle de ξ sachant ζ. Alors,

$$\begin{aligned} L(y_t \mid y_{t-1}, \delta_t = 0, \delta_{t-1} = 1) &= \frac{L(y_t, y_{t-1}, \delta_t = 0, \delta_{t-1} = 1)}{L(y_{t-1}, \delta_t = 0, \delta_{t-1} = 1)} \\ &= \frac{L(\delta_t = 0 \mid y_t, y_{t-1}, \delta_{t-1} = 1) L(y_t, y_{t-1}, \delta_{t-1} = 1)}{L(\delta_t = 0 \mid y_{t-1}, \delta_{t-1} = 1) L(y_{t-1}, \delta_{t-1} = 1)} \\ &= L(y_t \mid y_{t-1}, \delta_{t-1} = 1), \end{aligned}$$

où la troisième égalité découle de (1). De même, nous pouvons montrer que

$$L(y_t \mid y_{t-1}, \delta_t = 1, \delta_{t-1} = 1) = L(y_t \mid y_{t-1}, \delta_{t-1} = 1).$$

Donc, $L(y_t \mid y_{t-1}, \delta_t = 0, \delta_{t-1} = 1) = L(y_t \mid y_{t-1}, \delta_t = 1, \delta_{t-1} = 1)$ et le résultat (5) s'ensuit.

Preuve de (7)

En utilisant la même notation que dans la preuve précédente, nous avons

$$\begin{aligned} L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+1} = 0, \delta_r = 1) &= \frac{L(y_t, y_r, \delta_t = \cdots = \delta_{r+1} = 0, \delta_r = 1)}{L(y_r, \delta_t = \cdots = \delta_{r+1} = 0, \delta_r = 1)} \\ &= \frac{L(\delta_{r+1} = 0 \mid y_t, y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1)}{L(\delta_{r+1} = 0 \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1)} \\ &\quad \times \frac{L(y_t, y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1)}{L(y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1)} \\ &= L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1), \end{aligned}$$

où la dernière égalité découle de (1). De même, nous pouvons montrer que

$$\begin{aligned} L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1) \\ = L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_r = 1). \end{aligned}$$

Donc,

$$\begin{aligned} L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+1} = 0, \delta_r = 1) \\ = L(y_t \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1) \end{aligned}$$

et le résultat (7) s'ensuit.

Preuve du théorème 1

Posons que $t(= 2, \ldots, T)$ est fixe et que $n_{t,r} =$ le nombre d'unités avec $I_{t,r,i} = 1, r = 1, \ldots, t-1$. Nous commençons par montrer que, pour $r < t-1$,

$$E(\phi_{t,r+1}(y_{r+1}) \mid y_r, I_{t,r} = 1) = E(y_t \mid y_r, I_{t,r} = 1). \quad (11)$$

Sous l'hypothèse (1),

$$\begin{aligned} &L(y_t \mid y_{t-1}, \ldots, y_1, \delta_{t-1}, \ldots, \delta_1) \\ &= \frac{L(\delta_{t-1} \mid y_t, y_{t-1}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1)}{L(\delta_{t-1} \mid y_{t-1}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1)} L(y_t \mid y_{t-1}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1) \\ &= L(y_t \mid y_{t-1}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1) \\ &\cdots \\ &= L(y_t \mid y_{t-1}, \ldots, y_1) \\ &= L(y_t \mid y_{t-1}), \end{aligned}$$

où la dernière égalité découle de l'hypothèse (2). Alors

$$\begin{aligned} L(y_t \mid y_{t-1}, \delta_{t-1}) &= \int L(y_t \mid y_{t-1}, \ldots, y_1, \delta_{t-1}, \ldots, \delta_1) \\ &\quad dL(y_{t-2}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1 \mid y_{t-1}, \delta_{t-1}) \\ &= L(y_t \mid y_{t-1}) \int dL(y_{t-2}, \ldots, y_1, \delta_{t-2}, \ldots, \delta_1 \mid y_{t-1}, \delta_{t-1}) \\ &= L(y_t \mid y_{t-1}) \end{aligned}$$

et

$$\begin{aligned} L(y_t, \delta_t \mid y_{t-1}, \ldots, y_1, \delta_{t-1}, \ldots, \delta_1) &= L(\delta_t \mid y_t, y_{t-1}, \ldots, y_1, \delta_{t-1}, \ldots, \delta_1) \\ &\quad \times L(y_t \mid y_{t-1}, \ldots, y_1, \delta_{t-1}, \ldots, \delta_1) \\ &= L(\delta_t \mid y_{t-1}) L(y_t \mid y_{t-1}) \\ &= L(\delta_t \mid y_t, y_{t-1}, \delta_{t-1}) L(y_t \mid y_{t-1}, \delta_{t-1}) \\ &= L(y_t, \delta_t \mid y_{t-1}, \delta_{t-1}). \end{aligned}$$

Donc, $\{(y_t, \delta_t), t = 1, \ldots, T\}$ est une chaîne de Markov. Par conséquent,

$$\begin{aligned} E(y_t \mid y_{r+1}, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = 1) \\ = E(y_t \mid y_{r+1}, y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1). \end{aligned}$$

Alors, le premier membre de (11) est égal à

$$\begin{aligned} E[E(y_t \mid y_{r+1}, y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1) \mid \\ y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1] \\ = E(y_t \mid y_r, \delta_t = \cdots = \delta_{r+2} = 0, \delta_{r+1} = \delta_r = 1), \end{aligned}$$

qui est le deuxième membre de (11).

Il découle de la construction de $\tilde{y}_{t,i}$ décrite à la section 3, du résultat (11) et de la preuve du théorème 2.1 dans Cheng (1994) que

$$\sqrt{n_{t,r}}(\bar{y}_{t,r} - \mu_{t,r}) \to_d N(0, \sigma^2_{t,r})$$

pour une certaine $\sigma_{t,r}^2$, où $\bar{y}_{t,r} = 1/n_{t,r} \sum_{i=1}^{n} I_{t,r,i}\, \tilde{y}_{t,i}$ et $\mu_{t,r} = E[\phi_{t,r}(y_r) \,|\, I_{t,r} = 1], r = 1, \ldots, t-1$. Le résultat découle de

$$\begin{aligned}
\mu_t &= \sum_{r=1}^{t-1} E(y_t \,|\, I_{t,r} = 1) P(I_{t,r} = 1) \\
&= \sum_{r=1}^{t-1} E[E(y_t \,|\, y_r, I_{t,r} = 1) | I_{t,r} = 1] P(I_{t,r} = 1) \\
&= \sum_{r=1}^{t-1} E[\phi_{t,r}(y_r) | I_{t,r} = 1]\, P(I_{t,r} = 1) \\
&= \sum_{r=1}^{t-1} \mu_{t,r} P(I_{t,r} = 1)
\end{aligned}$$

et de

$$\begin{aligned}
&\sqrt{n}(\hat{\bar{Y}}_t - \mu_t) \\
&= \sum_{r=1}^{t-1} \left[\frac{\sqrt{n_{t,r}}}{\sqrt{n}} \sqrt{n_{t,r}} (\bar{y}_{t,r} - \mu_{t,r}) + \mu_{t,r} \sqrt{n} \left(P(I_{t,r} = 1) - \frac{n_{t,r}}{n} \right) \right].
\end{aligned}$$

Remerciements

Les auteurs remercient deux examinateurs et un rédacteur associé de leurs suggestions et commentaires constructifs. L'étude a été financée en partie par les subventions DMS-0404535 et SES-0705033 de la NSF.

Bibliographie

Bickel, P.J., et Freedman, D.A. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 470-482.

Butani, S., Harter, R. et Wolter, K. (1997). Estimation procedures for the Bureau of Labor Statistics Current Employment Statistics Program. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 523-528.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 81-87.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, Royaume Uni.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Krewski, D., et Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 1010-1019.

Little, R.J., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 1320-1329.

Robins, J.M., Rotnitzky, A. et Zhao, L.P. (1995). Analysis of semi-parametric regression method for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York : John Wiley & Sons, Inc.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. New York : CRC Press.

Shao, J., et Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 1278-1288.

Shao, J., et Tu, D. (1995). The Jackknife and Bootstrap. New York : Springer.

Troxel, A.B., Harrington, D.P. et Lipsitz, S.R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics*, 47, 425-438.

Troxel, A.B., Lipsitz, S.R. et Brennam, T.A. (1997). Weighted estimating equations with nonignorable missing response data. *Biometrics*, 53, 857-869.

Troxel, A.B., Lipsitz, S.R. et Harrigton, D.P. (1998). Marginal models for the analysis of longitudinal measurements with non-ignorable non-monotome missing data. *Biometrika*, 85, 661-672.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 89-96.

Xu, J. (2007). Methods for intermittent missing responses in longitudinal data. Thèse de doctorat, Department of Statistics, University of Wisconsin-Madison.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 181-192
Statistique Canada, N° 12-001-X au catalogue

Calage adaptatif pour la prédiction de totaux de population finie

Robert G. Clark et Raymond L. Chambers [1]

Résumé

Les poids d'échantillonnage peuvent être calés de manière à refléter les totaux connus de population d'un ensemble de variables auxiliaires. Le biais des prédicteurs des totaux de population finie calculés en utilisant ces poids est faible si ces variables sont reliées à la variable d'intérêt, mais leur variance peut être élevée si l'on utilise un trop grand nombre de variables auxiliaires. Dans le présent article, nous élaborons une approche de « calage adaptatif » où les variables auxiliaires qu'il convient d'utiliser dans la pondération sont sélectionnées en se servant de données d'échantillon. Nous montrons que, dans de nombreux cas, les estimateurs calés adaptativement ont une erreur quadratique moyenne plus faible et de meilleures propriétés de couverture que les estimateurs non adaptatifs.

Mots clés : Enquête par sondage ; pondération de l'échantillon ; approche de prédiction ; estimation ridge ; sélection du modèle ; méthodes séquentielles (pas à pas).

1. Introduction

Les prédicteurs de totaux de population finie sont habituellement calculés par sommation pondérée de valeurs d'échantillon. Très souvent, on dispose de variables auxiliaires dont les valeurs d'échantillon et les totaux de population sont connus. Les poids peuvent être construits de manière à ce que les sommes pondérées des variables auxiliaires concordent avec les totaux connus de population, processus appelé calage (Deville et Särndal 1992). Les prédicteurs des totaux de population finie fondés sur les poids calés ont généralement un biais de prédiction beaucoup plus faible que ceux calculés sans information auxiliaire.

Les auteurs de la littérature existante sur la prédiction en population finie supposent essentiellement qu'un ensemble de variables auxiliaires utiles est choisi sans référence aux données d'échantillon. Cependant, en pratique, il existe parfois un grand ensemble de variables auxiliaires possibles qui ne devraient pas toutes être utilisées. L'augmentation du nombre de variables auxiliaires réduit généralement le biais des prédicteurs calés, mais accroît la variance, de sorte qu'utiliser trop de variables auxiliaires peut effectivement augmenter l'erreur quadratique moyenne des prédicteurs calés. Souvent, le choix des variables auxiliaires qu'il convient d'utiliser n'est pas évident, et des données d'échantillon peuvent être nécessaires pour déterminer quel ensemble de ces variables est approprié pour les prédicteurs des totaux de variables d'intérêt particulières. Dans le présent article, nous élaborons des méthodes permettant de faire cette détermination. Notre approche peut être appelée calage adaptatif, parce que l'ensemble de variables est choisi adaptativement en partant des données d'échantillon, plutôt que statiquement, sans tenir compte de l'échantillon dont on dispose.

Nous utilisons le cadre de prédiction pour l'estimation en population finie (voir, par exemple, Brewer 1963 ; Royall 1970 ; Valliant, Dorfman et Royall 2000). Suivant cette approche, les valeurs de population des variables d'intérêt sont traitées comme des variables aléatoires. Le but est de prédire le total de population (qui est aussi une variable aléatoire) ou d'autres grandeurs de population finie en utilisant des données d'échantillon sur la variable d'intérêt, et des données de population sur certaines variables auxiliaires. L'échantillon peut avoir été sélectionné par échantillonnage probabiliste ou par une autre méthode, et est traité conditionnellement à la méthode choisie dans l'inférence. L'une des caractéristiques principales de l'approche est l'utilisation d'un modèle stochastique de la variable d'intérêt. L'un des aspects du cadre de prédiction est que la spécification incorrecte du modèle, par exemple par omission de variables auxiliaires importantes, peut causer un biais important.

Un autre cadre est celui de l'approche assistée par modèle (Särndal, Swensson et Wretman 1992). Dans cette approche, on utilise encore un modèle stochastique, mais celui-ci joue un rôle moins crucial. La nature aléatoire de l'échantillonnage est exploitée pour faire en sorte que les estimateurs soient approximativement sans biais, même si le modèle est incorrect. Quand le modèle est correct, les deux approches donnent des estimateurs approximativement sans biais, mais celle fondée sur un modèle produit généralement des estimateurs d'intérêt dont la variance est plus faible. Si le modèle est spécifié incorrectement, les prédicteurs fondés sur un modèle et les estimateurs de leur variance peuvent présenter un biais plus important, mais des méthodes fondées sur un modèle robuste ont été élaborées pour contourner ce problème. Par exemple, Royall et Herson (1973a, 1973b) discutent de la prédiction robuste, et Royall

1. Robert G. Clark et Raymond L. Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australie. Courriel : Robert_Clark@uow.edu.au.

et Cumberland (1978, 1981a, 1981b) élaborent des estimateurs de variance robustes à l'hétéroscédasticité. Pour des comparaisons du cadre de prédiction et du cadre assisté par modèle, voir, par exemple, Smith (1976) et Hansen, Madow et Tepping (1983).

Le problème de la sélection d'un ensemble de variables auxiliaires dans le cadre assisté par modèle a été étudié par Silva et Skinner (1997), ainsi que par Skinner et Silva (1997). Ils ont constaté que l'ajout de variables de calage réduit l'erreur quadratique moyenne (EQM) jusqu'à un certain point, après lequel l'addition d'autres variables accroît l'EQM. Choisir les variables de calage adaptativement, en se basant sur des données d'échantillon, ont donné de meilleures estimations que le calage sur toutes les variables ou sans en utiliser aucune. L'applicabilité de ces travaux à la prédiction fondée sur un modèle n'est pas claire, parce que le rôle joué par le modèle dans les deux cadres est très différent. Des modèles mal spécifiés peuvent donner lieu à des prédicteurs fondés sur le modèle considérablement biaisés, tandis que les estimateurs assistés par modèle sont approximativement sans biais, même si des variables importantes sont omises. Par conséquent, diverses stratégies de sélection du modèle pourraient être appropriées dans les deux cadres de travail. En outre, les différences entre diverses variantes devraient, en principe, être plus prononcées dans le cadre de prédiction que dans le cadre assisté par modèle.

Chambers, Skinner et Wang (1999) ont proposé une approche de sélection des variables de calage dans le cadre de prédiction qui s'appuie sur la sélection ascendante, descendante ou séquentielle. (Dans la suite du texte, cette approche sera appelée CSW.) Leur décision d'omettre (ou d'ajouter) une variable à chaque étape était fondée sur la minimisation de l'erreur quadratique estimée de la prédiction (EQMP) pour le prédicteur d'intérêt. L'approche n'a pas été évaluée par une étude en simulation, et les estimateurs de l'EQMP utilisés n'étaient pas robustes à l'hétéroscédasticité.

Le but du présent article est de développer l'approche de base de CSW afin de l'appliquer à une gamme plus étendue de situations, y compris les populations hétéroscédastiques et les échantillons à plusieurs degrés, et d'évaluer l'approche en utilisant des études par simulation réalistes. Nous utiliserons des estimateurs de l'EQMP qui sont robustes à l'hétéroscédasticité, ainsi qu'à la corrélation dans le cas des sondages à plusieurs degrés. Nous évaluerons les propriétés des estimateurs par simulation en nous servant de deux populations : des données financières sur les fermes provenant d'une enquête sur les fermes et des données sur la population active provenant d'un recensement de population.

Comme dans l'étude CSW, l'approche de base consistera à construire un ensemble de variables auxiliaires par sélection séquentielle de variables, en partant d'un ensemble initial de données. Cet algorithme construit un ensemble de variables auxiliaires grâce à une suite de décisions multiples de choix entre deux ensembles emboîtés de variables. Nous comparons plusieurs critères pour faire un choix entre les deux ensembles emboîtés, dont la signification statistique et un certain nombre d'estimateurs possibles de l'erreur quadratique moyenne de prédiction (EQMP). Nous considérons trois estimateurs différents de l'EQMP, à savoir un estimateur non robuste, un estimateur robuste à l'hétéroscédasticité et un estimateur robuste à l'hétéroscédasticité ainsi qu'aux corrélations entre les unités primaires d'échantillonnage dans l'échantillonnage à plusieurs degrés.

À la section 2, nous donnons la notation et les définitions. À la section 3, nous calculons la différence entre les EQMP de deux prédicteurs basés sur des modèles emboîtés et nous élaborons plusieurs options d'estimateur de cette différence. À la section 4, nous présentons les résultats des simulations pour une enquête sur les fermes et une enquête à plusieurs degrés auprès des ménages. La section 5 est consacrée à une discussion. Nous concluons que le calage adaptatif donne généralement de meilleurs résultats que le calage statique, à condition d'utiliser un estimateur non robuste de l'EQMP, ou la signification statistique, comme objectif dans la sélection du modèle.

2. Notation et définitions

Une variable d'intérêt Y_i est observée pour un échantillon s de n unités, qui est un sous-ensemble d'une population finie U contenant N unités. Le but est d'estimer le total de population $T_Y = \sum_{i \in U} Y_i$ et d'autres grandeurs en population finie de Y. Un vecteur de variables auxiliaires $\boldsymbol{x_i}$ est disponible pour $i = 1, ..., n$, dont le total de population $\boldsymbol{T}_x = \sum_{i \in U} \boldsymbol{x_i}$ est connu.

Les estimateurs pondérés de T_Y sont donnés par $\hat{T}_Y = \sum_{i \in s} w_i Y_i$, où w_i peut dépendre des variables auxiliaires, mais non de la variable d'intérêt. Un ensemble de poids est dit calé sur $\boldsymbol{x_i}$ si $\sum_{i \in s} w_i \boldsymbol{x_i} = \boldsymbol{T}_x$.

Le meilleur prédicteur linéaire sans biais (BLUP for *Best Linear Unbiased Predictor*) fondé sur un modèle de régression linéaire est un exemple d'estimateur calé. Le prédicteur BLUP utilisé le plus fréquemment est basé sur le modèle

$$E[Y_i] = \boldsymbol{\beta}^T \boldsymbol{x_i}$$

$$\text{var}[Y_i] = \sigma_i^2 = v_i \sigma^2 \qquad (1)$$

$$\text{cov}[Y_i, Y_j] = 0 (i \neq j)$$

(en supposant que v_i est connu) et est donné par

$$\hat{T}_Y = \sum_{i \in s} Y_i + \sum_{i \in r} \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i \qquad (2)$$

où $r = U - s$ est l'ensemble d'unités non incluses dans l'échantillon et

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i \in s} v_i^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^T \right\}^{-1} \sum_{i \in s} v_i^{-1} \boldsymbol{x}_i Y_i \qquad (3)$$

est un estimateur par les moindres carrés pondérés de $\boldsymbol{\beta}$. Le prédicteur BLUP peut aussi s'écrire sous la forme pondérée

$$\hat{T}_Y = \sum_{i \in s} w_i Y_i$$

où les poids w_i sont donnés par

$$w_i = 1 + \boldsymbol{T}_{\boldsymbol{xr}}^T \left\{ \sum_{j \in s} v_j^{-1} \boldsymbol{x}_j \boldsymbol{x}_j^T \right\}^{-1} v_i^{-1} \boldsymbol{x}_i \qquad (4)$$

et $\boldsymbol{T}_{\boldsymbol{xr}} = \sum_{i \in r} \boldsymbol{x}_i$. Il est facile de vérifier que $\sum_{i \in s} w_i \boldsymbol{x}_i = \boldsymbol{T}_{\boldsymbol{x}}$.

Dans le cas de données hétéroscédastiques, il est habituellement difficile de modéliser v_i de manière fiable. Dans ce cas, on dispose d'estimateurs robustes de la variance de prédiction du prédicteur BLUP qui ne requièrent pas que l'on connaisse v_i (Royall et Cumberland 1978, 1981a, 1981b). En ce qui concerne les échantillons à plusieurs degrés, l'hypothèse d'indépendance peut être violée. Le cas échéant, on peut encore utiliser le prédicteur BLUP basé sur (1) et recourir à un estimateur de variance robuste de la variance de prédiction basé sur les grappes finales (par exemple, Valliant et coll. 2000, chapitre 9). Une autre approche, que nous ne considérons pas ici, consisterait à construire un prédicteur BLUP basé sur un modèle dans lequel sont incluses les corrélations intra-grappe (Royall 1976). Nous discuterons plus en détail de l'estimation robuste et non robuste de l'erreur quadratique moyenne de la prédiction du BLUP à la section 3.

Nous devons prendre une décision quant aux éléments à inclure dans $\boldsymbol{x}_i$ dans le prédicteur BLUP. La sélection séquentielle, la sélection ascendante et la sélection descendante sont des algorithmes qui peuvent être utilisés pour décider quel sous-ensemble des variables auxiliaires disponibles il conviendrait d'utiliser. Les trois algorithmes comprennent chacun de nombreux choix entre deux ensembles emboîtés de variables auxiliaires. Supposons que nous devions choisir entre A) l'utilisation d'un prédicteur $\hat{T}_A$ basé sur $\boldsymbol{x}_i$ et B) l'utilisation d'un prédicteur $\hat{T}_B$ basé sur un sous-vecteur $\boldsymbol{x}_{1i}$. Nous pouvons partitionner $\boldsymbol{x}_i$ sous la forme $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}^T, \boldsymbol{x}_{2i}^T)^T$. Les nombres d'éléments de $\boldsymbol{x}_i$, $\boldsymbol{x}_{1i}$ et $\boldsymbol{x}_{2i}$ sont désignés par p, p_1 et p_2, respectivement.

Nous partitionnons de la même manière $\boldsymbol{\beta}$ sous la forme $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$. Le prédicteur $\hat{T}_A$ est sans biais sous le modèle A :

$$E[Y_i] = \boldsymbol{\beta}^T \boldsymbol{x}_i = \boldsymbol{\beta}_1^T \boldsymbol{x}_{1i} + \boldsymbol{\beta}_2^T \boldsymbol{x}_{2i}. \qquad (5)$$

Le prédicteur $\hat{T}_B$ est sans biais sous le modèle B,

$$E[Y_i] = \boldsymbol{\beta}_1^T \boldsymbol{x}_{1i}, \qquad (6)$$

qui est le cas particulier du modèle A où $\boldsymbol{\beta}_2 = \boldsymbol{0}$.

3. Estimation de la différence entre les EQMP

3.1 Comparaison des prédicteurs provenant de modèles hiérarchiques

À l'instar de CSW, notre approche consiste à estimer la différence entre les EQMP de deux estimateurs :

$$\Delta = E[(\hat{T}_A - T_Y)^2] - E[(\hat{T}_B - T_Y)^2]$$

où les espérances sont évaluées par rapport au modèle A, parce que le modèle B est un cas particulier de ce modèle. Habituellement, $\hat{T}_A$ aura un biais plus faible que $\hat{T}_B$, mais une variance plus élevée. L'un ou l'autre prédicteur peut avoir une EQMP plus élevée ou plus faible, selon la population et l'échantillon en question.

Dans le cas de l'échantillonnage à un seul degré, il est habituellement raisonnable de supposer que Y_i et Y_j sont indépendantes pour tout $i \neq j$. À la section 3.2, nous dérivons Δ et un estimateur de cette quantité dans ces conditions. À la section 3.3, nous décrivons le cas particulier instructif où les variances sont égales et où des prédicteurs BLUP sont utilisés ; il s'agit du cas pris en considération par CSW. À la section 3.4, nous généralisons ce cas particulier en décrivant un estimateur de Δ robuste à l'hétéroscédasticité. À la section 3.5, nous étendons encore davantage l'approche en calculant Δ et un estimateur de cette quantité sous échantillonnage à plusieurs degrés, où il peut exister des corrélations entre des valeurs provenant d'une même grappe.

3.2 Estimation de Δ dans un échantillonnage à un degré avec variance connue

En plus du modèle (5), nous supposons ici que Y_i et Y_j sont indépendantes pour $i \neq j$ et que $\text{var}[Y_i] = \sigma_i^2 = \sigma^2 v_i$ où v_i est connu. Dans ces conditions, l'EQMP de tout prédicteur $\hat{T} = \sum_{i \in s} w_i Y_i$ est donnée par

$$\begin{aligned} \text{EQMP}[\hat{T}] &= E[(\hat{T} - T_Y)^2] \\ &= \left\{ E\left[\sum_{i \in s} w_i Y_i - \sum_{i \in U} Y_i \right] \right\}^2 + \text{var}\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i \right] \\ &= \left\{ \boldsymbol{\beta}^T \left(\sum_{i \in s} w_i \boldsymbol{x}_i - \sum_{i \in U} \boldsymbol{x}_i \right) \right\}^2 + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2. \end{aligned}$$

En écrivant $\boldsymbol{d} = \sum_{i\in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$, nous pouvons réécrire l'EQMP sous la forme

$$\text{EQMP}[\hat{T}] = \boldsymbol{d}^T(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d} + \sum_{i\in s}(w_i - 1)^2\sigma_i^2 + \sum_{i\in r}\sigma_i^2.$$

Soit $\boldsymbol{d_A} = \sum_{i\in s} w_{Ai}\boldsymbol{x_i} - \boldsymbol{T_x}$ et $\boldsymbol{d_B} = \sum_{i\in s} w_{Bi}\boldsymbol{x_i} - \boldsymbol{T_x}$. Alors, Δ est donné par :

$$\begin{aligned}\Delta &= \text{EQMP}[\hat{T}_A] - \text{EQMP}[\hat{T}_B] \\ &= \boldsymbol{d_A^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_A} - \boldsymbol{d_B^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\sigma_i^2 - \sum_{i\in s}(w_{Bi} - 1)^2\sigma_i^2. \qquad (7)\end{aligned}$$

Pour estimer Δ, nous considérons d'abord la façon d'estimer $\boldsymbol{\beta}$ et la variance de $\hat{\boldsymbol{\beta}}$. L'estimateur par les moindres carrés habituels est $\hat{\boldsymbol{\beta}} = S_x^{-1}S_{xy}$ où $S_x = \sum_{i\in s} v_i^{-1}\boldsymbol{x_i}\boldsymbol{x_i^T}$ et $S_{xy} = \sum_{i\in s} v_i^{-1}\boldsymbol{x_i}Y_i$. L'estimateur habituel de la variance de $\hat{\boldsymbol{\beta}}$ est $\widehat{\text{var}}[\hat{\boldsymbol{\beta}}] = \hat{\sigma}^2 S_x^{-1}$, où $\hat{\sigma}^2 = \sum_{i\in s}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2 v_i^{-1}/(n-p)$.

Nous pouvons estimer $(\boldsymbol{\beta}\boldsymbol{\beta}^T)$ sans biais par $(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])$. Par conséquent, l'expression suivante est un estimateur sans biais de Δ :

$$\begin{aligned}\hat{\Delta} &= \boldsymbol{d_A^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_A} - \boldsymbol{d_B^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\hat{\sigma}^2 v_i - \sum_{i\in s}(w_{Bi} - 1)^2\hat{\sigma}^2 v_i. \qquad (8)\end{aligned}$$

L'expression (7) est applicable et l'estimateur (8) en est un estimateur sans biais, pour tout prédicteur pondéré $\hat{T}_A$ et $\hat{T}_B$. Nous nous intéressons au cas particulier où $\hat{T}_A$ et $\hat{T}_B$ sont des prédicteurs BLUP. Dans ce cas, $\hat{T}_A$ est calé sur $\boldsymbol{T_x}$ de sorte que $\boldsymbol{d_A} = \boldsymbol{0}$, et (8) se simplifie alors en

$$\begin{aligned}\hat{\Delta} &= -\boldsymbol{d_B^T}(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai} - 1)^2\hat{\sigma}^2 v_i - \sum_{i\in s}(w_{Bi} - 1)^2\hat{\sigma}^2 v_i. \qquad (9)\end{aligned}$$

3.3 Un cas particulier important

À la présente sous-section, nous faisons les hypothèses énoncées à la section 3.2 et nous supposons en outre que $v_i = 1$ pour tout i. Nous supposons aussi que la dimension de $\boldsymbol{x_{2i}}$ est 1, autrement dit nous considérons s'il faut utiliser ou non dans la prédiction une variable auxiliaire particulière provenant de $\boldsymbol{x_i}$. Les expressions (7) et (9) se simplifient dans ces conditions.

Soit u_i le résidu d'une régression de x_{2i} sur $\boldsymbol{x_{1i}}$:

$$u_i = x_{2i} - C^T\boldsymbol{x_{1i}}$$

$$C = \left(\sum_{i\in s}\boldsymbol{x_{1i}}\boldsymbol{x_{1i}^T}\right)^{-1}\sum_{i\in s}\boldsymbol{x_{1i}}x_{2i}.$$

À l'aide d'opérations algébriques linéaires simples, nous pouvons montrer que :

$$\boldsymbol{\beta}^T\boldsymbol{d_B} = -\beta_2\sum_{i\in r}u_i$$

et que

$$\sum_{i\in s}(w_{Ai} - 1)^2 - \sum_{i\in s}(w_{Bi} - 1)^2 = \left(\sum_{i\in r}u_i\right)^2 S_u^{-1}$$

où $S_u = \sum_{i\in s}u_i^2$.

Donc, (7) devient

$$\Delta = \sigma^2\left(\sum_{i\in r}u_i\right)^2 S_u^{-1} - \left(\sum_{i\in r}u_i\right)^2\beta_2^2.$$

CSW montrent que $d_B^T\,\widehat{\text{var}}[\hat{\beta}_2]\,d_B = \hat{\sigma}^2(\sum_{i\in r}u_i)^2 S_u^{-1}$. Par conséquent, (9) devient

$$\hat{\Delta} = \left(\sum_{i\in r}u_i\right)^2(2\hat{\sigma}^2 S_u^{-1} - \hat{\beta}_2^2).$$

Il est proposé d'adopter $\hat{T}_A$ quand $\hat{\Delta} < 0$, et d'utiliser $\hat{T}_B$ autrement. Il s'ensuit que nous adoptons $\hat{T}_A$ quand $\hat{\beta}_2^2 > 2\hat{\sigma}^2 S_u^{-1}$. Comme l'ont mentionné CSW, cela équivaut à adopter $\hat{T}_A$ quand $F = \hat{\beta}_2^2/(\hat{\sigma}^2 S_u^{-1})$ est supérieur à 2. Notons que F est la statistique utilisée habituellement pour tester l'hypothèse nulle selon laquelle $\beta_2 = 0$. Si n est grand, le seuil pour le test F au niveau de signification de 5 % est 3,96, tandis que nous sommes arrivés à un seuil de 2 pour l'adoption de l'ensemble plus grand de variables. Par conséquent, la décision d'utiliser A au lieu de B en s'appuyant sur un test de signification requiert plus de preuve à l'encontre de B que ne le laisserait supposer une simple comparaison des EQMP estimées de $\hat{T}_A$ et $\hat{T}_B$. Autrement dit, utiliser $\hat{\Delta}$ donne lieu à de plus grands modèles que l'utilisation d'un test de signification.

3.4 Estimation de Δ robuste à l'hétéroscédasticité

Les estimateurs de Δ présentés aux sections 3.2 et 3.3 s'appuient sur le fait que $\text{var}[Y_i]$ est connue au moins jusqu'à une constante de proportionnalité. En pratique, les variances sont, au mieux, connues approximativement et les méthodes qui ne dépendent pas de l'hypothèse que la variance est connue pourraient donner de meilleurs résultats. Nous utiliserons un estimateur de σ_i^2 qui, selon l'hypothèse du modèle (5), est approximativement sans biais pour σ_i^2 en général, et exactement sans biais si $\sigma_i^2 = \sigma^2$:

$$\hat{\sigma}_i^2 = \frac{n}{n-p}(Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2.$$

(Un estimateur de rechange serait $\hat{\sigma}_i^2 = (Y_i - \hat{\boldsymbol{\beta}}^T\boldsymbol{x_i})^2$, comme dans Royall et Cumberland 1981b.)

L'estimateur de $\boldsymbol{\beta}$ sera encore l'estimateur par les moindres carrés pondérés donné par (3). La variance de $\hat{\boldsymbol{\beta}}$ est

$$\begin{aligned}\text{var}[\hat{\boldsymbol{\beta}}] &= \text{var}[S_x^{-1} S_{xy}] \\ &= \text{var}\left[S_x^{-1} \sum_{i \in s} \boldsymbol{x_i} Y_i\right] \\ &= S_x^{-1}\left(\sum_{i \in s} \boldsymbol{x_i}\, \boldsymbol{x_i^T}\, \sigma_i^2\right) S_x^{-1}.\end{aligned}$$

Elle peut être estimée par

$$\widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}] = S_x^{-1}\left(\sum_{i \in s} \boldsymbol{x_i}\, \boldsymbol{x_i^T}\, \hat{\sigma}_i^2\right) S_x^{-1}.$$

Donc, nous pouvons estimer Δ par

$$\begin{aligned}\hat{\Delta} &= \boldsymbol{d_A^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - \widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d_A} \\ &\quad - \boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - \widehat{\text{var}}_{\text{robust}}[\hat{\boldsymbol{\beta}}]\right) \boldsymbol{d_B} \\ &\quad + \sum_{i \in s} (w_{Ai} - 1)^2\, \hat{\sigma}_i^2 - \sum_{i \in s} (w_{Bi} - 1)^2\, \hat{\sigma}_i^2. \qquad (10)\end{aligned}$$

3.5 Estimation de Δ dans l'échantillonnage à plusieurs degrés

Les estimateurs de Δ présentés aux sections 3.2, 3.3 et 3.4 reposent tous sur l'hypothèse que les valeurs de Y sont indépendantes pour différentes unités. Dans l'échantillonnage à plusieurs degrés, on commence par sélectionner un échantillon d'unités primaires d'échantillonnage (UPE), puis on tire un échantillon d'unités dans les UPE sélectionnées. Par exemple, les UPE pourraient être des régions et les unités, des ménages ou des personnes ; ou bien les UPE pourraient être des écoles et les unités, des élèves. Habituellement, les unités provenant d'une même UPE ont tendance à être semblables, de sorte que les valeurs de Y_i et Y_j peuvent être corrélées si i et j appartiennent à la même UPE. À la présente section, nous élaborons un estimateur de Δ qui est approximativement sans biais, même quand il existe des corrélations entre les valeurs de Y dans une même UPE.

Soit s_I l'échantillon d'UPE, tiré de la population U_I. Soit s_g l'échantillon d'unités tirées de l'UPE g, où $g \in s_I$. Soit $r_I = U_I - s_I$ et $r_g = U_g - s_g$. Nous émettons l'hypothèse du modèle (5) et supposons en outre que Y_i et Y_j ne sont pas corrélées pour $i \in g_1$ et $j \in g_2$ si $g_1 \neq g_2$. Les valeurs de Y_i et Y_j peuvent être corrélées si $i \neq j$ avec $i, j \in U_g$.

Soit $\hat{T} = \sum_{i \in s} w_i Y_i$ tout prédicteur et soit $\boldsymbol{d} = \sum_{i \in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$. Le biais de $\hat{T}$ est $\boldsymbol{\beta^T d}$, comme à la section 3.2. La variance de $(\hat{T} - T_Y)$ est

$$\begin{aligned}\text{var}[\hat{T} - T_Y] &= \text{var}\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i\right] \\ &= \text{var}\left[\sum_{g \in s_I}\left(\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right) - \sum_{g \in r_I} \sum_{i \in U_g} Y_i\right] \\ &= \sum_{g \in s_I} \text{var}\left(\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right) + \sum_{g \in r_I} \text{var}\left[\sum_{i \in U_g} Y_i\right].\end{aligned}$$

Nous supposons de surcroît que la variance de $\sum_{i \in r_g} Y_i$ et la covariance entre $\sum_{i \in r_g} Y_i$ et $\sum_{i \in s_g} (w_i - 1) Y_i$ sont négligeables comparativement aux autres termes. Il en est ainsi quant on recourt à l'échantillonnage en grappes (parce que, dans ce cas, $s_g = U_g$ and r_g est vide) ou que la fraction d'échantillonnage dans les UPE est faible. La variance devient

$$\text{var}\,[\hat{T} - T_Y] \approx \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_i - 1) Y_i\right] + \sum_{g \in r_I} \text{var}\left[\sum_{i \in U_g} Y_i\right].$$

En appliquant cela à Δ, nous obtenons :

$$\begin{aligned}\Delta &= \text{EQMP}[\hat{T}_A] - \text{EQMP}[\hat{T}_B] \\ &= \boldsymbol{d_A^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_A} - \boldsymbol{d_B^T}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\boldsymbol{d_B} + \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_{Ai} - 1) Y_i\right] \\ &\quad - \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} (w_{Bi} - 1) Y_i\right]. \qquad (11)\end{aligned}$$

Pour estimer Δ, nous avons besoin des estimateurs de la variance de $\hat{\boldsymbol{\beta}}$ et de $(\sum_{i \in s_g} (w_i - 1) Y_i)$.

En premier lieu, notons que

$$\begin{aligned}\text{var}\,[\hat{\boldsymbol{\beta}}] &= \text{var}\left[S_x^{-1} \sum_{g \in s_I} \sum_{i \in s_g} \boldsymbol{x_i} Y_i\right] \\ &= S_x^{-1} \sum_{g \in s_I} \text{var}\left[\sum_{i \in s_g} \boldsymbol{x_i} Y_i\right] S_x^{-1}.\end{aligned}$$

Nous pouvons estimer cette variance en utilisant la méthode de la variance des grappes finales (*ultimate cluster variance ou UCV*) selon

$$\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}] = S_x^{-1} \sum_{g \in s_I}\left(\sum_{i \in s_g} \boldsymbol{x_i}(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i})\right)^2 S_x^{-1}.$$

Cet estimateur bien connu de la variance d'une somme pondérée de données en grappes est équivalent à celui donné par Valliant et coll. (2000, 9.5.5, page 312). Cet estimateur a été appelé « estimateur sandwich de la variance en utilisant les résidus au niveau de la grappe » (Valliant et coll. 2000) et « de la variance des grappes finales » (par exemple, Wolter 1985 décrit essentiellement la même idée dans un cadre de randomisation).

La variance de $(\sum_{i \in s_g} (w_i - 1) Y_i)$ peut aussi être estimée par la méthode d'estimation de la variance des grappes finales :

$$\widehat{\text{var}}\left[\sum_{i\in s_g}(w_i-1)Y_i\right]=\left\{\sum_{i\in s_g}(w_i-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2.$$

Donc, nous pouvons estimer Δ par

$$\begin{aligned}\hat{\Delta}_{\text{ucv}} &= \boldsymbol{d}_A^T(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T-\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d}_A\\ &\quad-\boldsymbol{d}_B^T(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T-\widehat{\text{var}}_{\text{ucv}}[\hat{\boldsymbol{\beta}}])\,\boldsymbol{d}_B\\ &\quad+\sum_{g\in s_I}\left\{\sum_{i\in s_g}(w_{Ai}-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2\\ &\quad-\sum_{g\in s_I}\left\{\sum_{i\in s_g}(w_{Bi}-1)(Y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)\right\}^2. \qquad (12)\end{aligned}$$

4. Étude par simulation

4.1 Simulation d'une enquête sur les fermes

Population et plan d'échantillonnage

La distribution de population des variables auxiliaires, la taille d'échantillon et la taille de population, ainsi que l'hétéroscédasticité et d'autres propriétés de la variable d'intérêt sont des éléments qui, en principe, devraient tous jouer un rôle dans la performance des prédicteurs BLUP adaptatifs. Afin d'évaluer raisonnablement les propriétés de ces estimateurs, il est nécessaire de procéder à une étude par simulation basée sur une grande population réaliste.

Nous avons généré une population simulée de 80 000 unités à l'aide de données d'échantillon sur 1 652 fermes provenant de l'Australian Agricultural and Grazing Industry Survey (AAGIS) de 1988 comme point de départ. Nous avons choisi comme variable d'intérêt le total des cultures commerciales et avons considéré comme variables auxiliaires possibles l'EDT (estimation dérivée de la taille), le nombre de moutons, la superficie cultivée, le nombre de bovins, la région (29 régions) et l'industrie (5 industries). L'EDT était une combinaison linéaire des variables nombre de moutons, superficie cultivée et nombre de bovins. L'ensemble de données contenait aussi un poids d'échantillonnage qui était approximativement égal à l'inverse de la probabilité de sélection. Nous avons éliminé 27 cas aberrants pour lesquels la valeur de l'EDT était très grande, car, dans une enquête, ces cas seraient placés dans une strate entièrement dénombrée. Nous avons ensuite construit une population de 80 000 unités selon un plan d'échantillonnage avec probabilité proportionnelle à la taille avec remise, les probabilités étant proportionnelles au poids d'estimation figurant dans le fichier de l'échantillon original.

Ensuite, nous avons sélectionné 250 échantillons sans remise à partir de la population simulée. Nous avons stratifié les échantillons selon la région et l'EDT, en divisant l'EDT en quatre catégories, ce qui a donné 116 strates. Les limites des catégories ont été déterminées de telle façon que les sommes des catégories de l'EDT soient égales. Nous avons simulé des tailles totales d'échantillon de 250, 500, 1 000 et 1 500. Les tailles d'échantillon de strate étaient proportionnelles aux tailles d'échantillon de l'AAGIS originale selon la région et l'EDT.

Variables auxiliaires et méthode de sélection séquentielle

Nous avons inclus les variables auxiliaires correspondant au modèle contenant une constante (ordonnée à l'origine), le nombre de moutons (x1), la superficie cultivée (x2), le nombre de bovins (x3), l'industrie, l'interaction de l'industrie et de x1, x2 et x3, ainsi que la région. Cela donne un total de 52 variables auxiliaires possibles. Certaines de ces variables sont colinéaires, mais sont néanmoins incluses dans l'ensemble de variables possibles, afin de donner pour le processus de sélection du modèle un choix plus étendu de modèles possibles. Nous avons également considéré l'ensemble de 139 variables auxiliaires contenant cet ensemble, ainsi que l'interaction de la région et de x1, x2 et x3. Nous avons construit les modèles par sélection ascendante, en partant du modèle contenant uniquement la constante. Nous avons ajouté les variables en prenant pour critère l'étape qui réduisait le plus l'EQMP estimée, pour divers estimateurs possibles de Δ. Nous avons également essayé la sélection séquentielle, mais son exécution était considérablement plus lente pour le plus grand ensemble de variables et son utilisation n'a pas amélioré fortement l'efficacité des prédicteurs BLUP adaptatifs.

Nous avons également calculé un prédicteur BLUP adaptatif basé sur le test de signification statistique, en prenant $p < 0,05$ comme seuil d'inclusion. Pour chaque modèle consécutif, nous avons déterminé la signification statistique de chacune des variables non incluses dans le modèle à l'aide d'un test t standard. À chaque étape, nous avons inclus dans le modèle la variable dont la valeur p était la plus faible. Quand il n'a plus été possible d'ajouter d'autres variables ayant un effet significatif, nous avons arrêté la procédure et le modèle résultant est celui qui a été choisi.

Un certain nombre de modifications ont été nécessaires pour que l'algorithme de sélection ascendante fonctionne fiablement : les variables auxiliaires n'ont pas été ajoutées au modèle si la corrélation de paire de Pearson était égale ou supérieure à 0,95 (ou égale ou inférieure à -0,95) ; et les variables n'ont pas été ajoutées si leur ajout donnait lieu à des équations de calage impossibles à résoudre.

Estimateurs utilisés

Nous avons calculé plusieurs prédicteurs BLUP, en incluant toutes les variables auxiliaires, en incluant

uniquement la constante et l'EDT, et en incluant les variables auxiliaires choisies par sélection ascendante en utilisant l'estimateur non robuste de Δ (décrit à la section 3.2) ou l'estimateur robuste à l'hétéroscédasticité de Δ (décrit à la section 3.4), en partant de l'ensemble de 52 et de l'ensemble de 139 variables auxiliaires possibles. (Le grand ensemble de 139 variables a été évalué uniquement pour des tailles d'échantillon égales ou supérieures à 500.)

L'utilisation d'estimateurs ridge (par exemple, Bardsley et Chambers 1984) est un autre moyen d'aborder le problème de sélection des variables, si bien que nous les avons inclus dans la simulation pour comparer leurs propriétés à celles des prédicteurs BLUP adaptatifs. Les estimateurs que nous avons pris en considération jusqu'à présent incluent ou excluent chaque variable. Si une variable est incluse, les poids doivent être calés exactement sur cette variable, en ce sens que $\sum_{i\in s} w_i \boldsymbol{x_i} = \boldsymbol{T_x}$. La régression ridge introduit une pénalité pour l'absence de calage, mais ne requiert pas nécessairement que les poids donnent un calage parfait pour toutes les variables. Dans cette régression, le vecteur de poids d'échantillon $\boldsymbol{w}$ est choisi de manière à minimiser

$$\sum_{i\in s} (w_i - 1)^2 v_i^{-1} + \sum_{j=1}^{p} c_j^{-1}\left(\sum_{i\in s} w_i x_{ij} - T_{xj}\right)^2.$$

Les c_j sont des coefficients de coût non négatifs indiquant avec quelle priorité il faut satisfaire la contrainte de calage j. Une valeur de 0 indique que la contrainte doit être satisfaite précisément, tandis que les coefficients de coût plus élevé accordent moins de poids à la contrainte. Donc, l'estimateur ridge permet une réduction lisse de la dimension effective du modèle, en procédant effectivement à une interpolation entre l'inclusion d'une variable de calage $(c_j = 0)$ et son exclusion $(c_j = \infty)$.

Habituellement, les c_j sont fixés à λc_j^*, où c_j^* reflète une évaluation un peu subjective de l'importance relative de chaque contrainte, et λ est choisi de manière que les poids finaux w_i aient des propriétés raisonnables, par exemple, qu'ils soient tous supérieurs ou égaux à 0, ou à 1. Nous fixons c_j^* à 0 pour la constante (reflétant une ordonnée à l'origine dans le modèle), à 1 pour $x1$, $x2$ et $x3$, à 10 pour les indicateurs de région, à 5 pour les indicateurs d'industrie, et à 100 pour les interactions. Nous avons choisi c_j^* en nous fondant sur les variables qui, à notre avis, étaient susceptibles d'être les plus utiles. Pour chaque échantillon, nous avons déterminé numériquement la valeur λ de façon qu'elle soit la valeur la plus faible permettant que tous les poids soient supérieurs ou égaux à 1.

Nous avons fondé toutes les méthodes sur la même procédure de modélisation de $\text{var}_M[Y_i]$. En premier lieu, nous avons ajusté un simple modèle contenant la constante, x1, x2 et x3 aux valeurs d'échantillon de Y en utilisant les moindres carrés ordinaires. Puis, nous avons procédé à la régression du logarithme des carrés des résidus de ce modèle sur le logarithme de l'EDT. Ensuite, nous avons élevé les valeurs ajustées de ce modèle à la puissance e pour obtenir les estimations de σ_i^2 pour chaque $i \in s$. Nous avons ensuite tronqué les valeurs estimées de σ_i^2 de sorte qu'aucune valeur ne soit supérieure ou inférieure d'un facteur 4 à la valeur médiane. Cet ajustement avait pour but d'éviter les valeurs extrêmes de σ_i^{-2} qui pourraient entraîner une instabilité dans le calcul des estimations par les moindres carrés pondérés de $\hat{\boldsymbol{\beta}}$. Les résultats étaient sensibles dans une certaine mesure à la méthode de modélisation de la variance, particulièrement l'ajustement final pour éviter les valeurs extrêmes. Les prédicteurs BLUP basés sur un modèle de variance grossier avec $\sigma_i^2 \propto \text{DSE}_i$ avaient une variance supérieure de 10 à 20 % à celle des prédicteurs BLUP présentés ici.

Résultats

Le tableau 1 donne la racine de l'erreur quadratique moyenne relative (RRMSE pour *Relative Root Mean Squared Error*) des divers prédicteurs calés. Les quatre premières lignes du tableau correspondent au premier ensemble de variables auxiliaires (52 variables possibles) et les trois dernières, au deuxième ensemble (139 variables auxiliaires possibles). Les biais ne sont pas présentés, mais constituaient généralement une composante relativement petite de l'erreur quadratique moyenne pour tous les prédicteurs présentés, sauf le prédicteur BLUP basé sur un modèle contenant la constante et l'EDT, qui était relativement biaisé. Ce résultat était quelque peu surprenant, car nous nous attendions à ce qu'un bon compromis entre le biais et la variance implique que les biais soient une composante non négligeable de l'erreur quadratique moyenne. Des détails sur les biais et les variances relatives des prédicteurs peuvent être consultés dans les tableaux A1 et A2 de Clark et Chambers (2008).

Parmi les prédicteurs BLUP adaptatifs, les critères de signification ont donné les meilleurs résultats dans tous les cas, suivis par les critères de non-robustesse, tandis que les critères de robustesse ont donné les pires résultats. Pour le petit ensemble de 52 variables possibles, les prédicteurs BLUP adaptatifs basés sur les critères de non-robustesse et les critères de signification ont donné de meilleurs résultats que les prédicteurs BLUP non adaptatifs pour $n = 250$ et $n = 500$; pour $n = 1\,000$ et $n = 1\,500$, ils ont donné d'un peu moins bons résultats que les prédicteurs BLUP contenant toutes les variables, mais de meilleurs résultats que le prédicteur BLUP contenant la constante et la variable de taille. Pour le grand ensemble de 139 variables possibles, les prédicteurs BLUP adaptatifs basés sur les critères de non-robustesse et de signification ont donné de meilleurs

résultats que les prédicteurs BLUP non adaptatifs pour toutes les tailles d'échantillon, particulièrement pour les faibles valeurs de n.

Tableau 1
RRMSE (%) des prédicteurs du total des cultures commerciales pour l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptif			Ridge
		Toutes	C^{ste} + taille	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.	
52	250	3,59	3,02	2,97	3,09	2,87	3,30
	500	2,35	2,54	2,33	2,33	2,30	2,31
	1 000	1,56	2,21	1,58	1,64	1,57	1,54
	1 500	1,36	2,22	1,39	1,41	1,37	1,37
139	500	3,52	2,54	2,99	3,44	2,29	2,27
	1 000	1,77	2,21	1,75	1,92	1,72	1,59
	1 500	1,56	2,22	1,51	1,64	1,42	1,42

En général, l'estimateur ridge a donné des résultats presque aussi bons que le meilleur des prédicteurs BLUP adaptatifs quand nous avons utilisé 52 variables auxiliaires et des résultats un peu meilleurs quand nous nous sommes servis de 139 variables auxiliaires possibles.

Le tableau 2 montre combien de variables auxiliaires ont été sélectionnées pour les deux prédicteurs BLUP adaptatifs. L'estimateur $\hat{\Delta}$ robuste a donné de plus grands ensembles de variables auxiliaires que l'estimateur non robuste, environ 10 variables auxiliaires supplémentaires ayant été sélectionnées. Les critères de signification ont produit des ensembles de variables encore plus petits (6 à 10 variables de moins que dans le cas des critères de non-robustesse).

Tableau 2
Moyenne (intervalle interquartile) du nombre de variables auxiliaires sélectionnées dans l'AAGIS

N^{bre} de var.	n	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.
52	250	16,0 (14,0-18,0)	26,9 (24,0-29,0)	9,6 (8,0-11,0)
	500	18,6 (16,0-21,0)	27,4 (25,0-30,0)	11,5 (10,0-13,0)
	1 000	23,6 (21,0-26,0)	29,6 (26,0-33,0)	14,4 (13,0-16,0)
	1 500	27,3 (25,0-29,0)	32,3 (30,0-35,0)	17,2 (16,0-18,8)
139	500	42,1 (37,0-47,0)	69,4 (62,0-75,0)	23,2 (21,0-26,0)
	1 000	51,5 (47,0-56,0)	74,2 (69,0-79,8)	29,9 (27,0-33,0)
	1 500	59,2 (55,0-64,0)	75,8 (71,0-81,0)	34,9 (32,0-38,0)

Le tableau 3 donne la non-couverture de l'intervalle de confiance (IC) des divers prédicteurs. Nous avons défini l'IC à 90 % comme étant l'estimateur +/- 1,64 erreur-type, où la variance était estimée en utilisant l'estimateur de variance robuste à l'hétéroscédasticité (Royall et Cumberland 1978). Conformément à une pratique courante, nous avons fondé les IC sur la variance estimée et non sur l'erreur quadratique moyenne estimée de prédiction. Les estimations par simulation des taux de non-couverture sont assez approximatives, étant donné que nous n'avons utilisé que 250 simulations. Une plus grande étude par simulation peut être effectuée pour obtenir des estimations plus précises de la couverture, mais nous n'avons pas poursuivi cette option étant donné que le processus de sélection séquentielle est gourmand en ressources informatiques. Le tableau 3 donne à penser que le prédicteur BLUP n'utilisant que la constante et la variable de taille donne une non-couverture importante, de même que le prédicteur BLUP adaptatif basé sur l'estimateur $\hat{\Delta}$ robuste. En général, le taux de non-couverture des autres estimateurs est proche des 10 % nominaux.

Tableau 3
Non-couverture de l'intervalle de confiance dans l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptif			Ridge
		Toutes	C^{ste} + taille	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	Test de sign.	
52	250	10,0	6,4	10,4	16,8	11,2	10,0
	500	8,0	13,2	12,0	17,2	10,8	8,0
	1 000	7,6	20,4	9,2	12,0	8,4	8,4
	1 500	8,8	34,8	9,2	13,2	9,6	8,8
139	500	16,8	13,2	18,0	29,2	12,8	8,8
	1 000	12,4	20,4	14,0	20,4	13,2	7,2
	1 500	13,6	34,8	13,6	19,6	12,4	11,2

Le total des cultures commerciales est une variable d'intérêt majeure de l'AAGISS, mais les totaux d'autres variables sont importants également, y compris les capitaux propres des fermes. Pour des raisons pratiques, un seul ensemble de poids est normalement utilisé pour toutes les variables. Le tableau 4 montre dans quelle mesure les poids de calage adaptatifs conçus pour le total des cultures commerciales (TCC) donnent de bons résultats quand on les utilise pour estimer le total des capitaux propres des fermes. Dans le cas où 52 variables auxiliaires possibles sont utilisées, les poids BLUP adaptatifs choisis en se fondant sur la variable TCC (en utilisant l'estimateur $\hat{\Delta}$ non robuste) ont donné d'assez bons résultats, de même que l'estimateur ridge. Cependant, des améliorations pourraient être apportées en choisissant les variables auxiliaires en ne se basant que sur la variable des capitaux propres.

Tableau 4
RRMSE (%) des prédicteurs du total des capitaux propres dans l'AAGIS

N^{bre} de var.	n	BLUP		BLUP adaptatif ($\hat{\Delta}$ non robuste)		Ridge
		Toutes	C^{ste} + taille	Basé sur TCC	Basé sur capitaux propres	
52	250	6,85	6,45	6,51	6,13	6,78
	500	4,44	4,44	4,61	4,40	4,28
	1 000	3,09	3,12	3,42	3,14	3,10
	1 500	2,54	2,58	2,90	2,58	2,54
139	500	5,53	4,93	4,98	4,74	4,20
	1 000	3,68	4,03	3,23	3,15	3,08
	1 500	3,04	3,63	2,66	2,60	2,57

4.2 Simulation de l'enquête sur la population active

Population et plan d'échantillonnage

Nous avons construit une population simulée en tirant un échantillon aléatoire simple sans remise de 30 000 personnes de 15 à 64 ans du fichier de l'échantillon à 1 % de l'Australian Census of Population and Housing de 1991. La variable d'intérêt était l'emploi (1 pour les personnes occupées, 0 pour les autres). Nous avons divisé la population simulée en unités primaires d'échantillonnage (UPE) simulées contenant 75 personnes chacune, de telle sorte que la corrélation intra-grappe soit égale à 0,05. (Il s'agit d'une valeur assez typique de la corrélation intra-grappe pour la variable d'emploi dans les unités primaires d'échantillonnage d'une enquête-ménage (voir, par exemple, Clark et Steel 2002). L'algorithme de définition des grappes devait trier les données en fonction d'une variable généré aléatoirement de loi $N(0, \gamma^2)$ plus la variable d'emploi, puis définir les grappes comme des ensembles séquentiels de 75 personnes, où γ était choisi de manière à obtenir la corrélation intra-grappe souhaitée.

La simulation consistait en 250 échantillons à deux degrés répétés. Le premier degré était un échantillon aléatoire simple sans remise de m UPE et le deuxième degré était un échantillon aléatoire simple sans remise de 20 personnes provenant de chaque UPE sélectionnée. La taille totale d'échantillon a été fixée à $n = 200, 400$ et 1 000 personnes. La plupart des enquêtes-ménages nationales sont réalisée auprès d'échantillons de taille beaucoup plus grande que cela, mais il est courant, pour l'estimation, de construire des poststrates dans les strates ou les provinces, et les tailles d'échantillon pour ces régions sont souvent de l'ordre de 200 à 1 000.

Les variables auxiliaires possibles étaient l'âge selon le sexe, où l'âge était enregistré en année simple pour le groupe des 16 à 24 ans, puis en tranche de cinq ans pour les groupes des 25 à 29 ans, 30 à 34 ans, ..., 55 à 59 ans et 60 ans et plus.

Non-réponse

L'une des raisons principales pour lesquelles on se sert de l'âge et du sexe comme variables auxiliaires dans les enquêtes-ménages est que l'on sait que la non-réponse dépend de ces variables. Par exemple, les jeunes hommes représentent habituellement le groupe dont le taux de réponse est le plus faible. Nous avons simulé la non-réponse en supposant que le logit de la probabilité de réponse était égal à $1{,}8 - ((\hat{a}ge - 50)/25)^2$ pour les hommes, et à $2 - 0{,}7((\hat{a}ge - 50)/25)^2$ pour les femmes. Ce modèle donne un taux de réponse de 75 %. Nous avons augmenté la taille de l'échantillon initial de manière que la taille de l'échantillon final de répondants soit égale à $n = 200, 400$ ou 1 000.

Variables auxiliaires et méthode de sélection séquentielle

Nous avons choisi les variables auxiliaires possibles en nous basant sur les cellules déterminées par l'âge selon le sexe. La définition des variables x est donnée au tableau 5. Nous avons choisi cette paramétrisation afin que les variables auxiliaires correspondant à des âges ou groupes d'âge particuliers puissent être abandonnées tout en obtenant encore un modèle raisonnable. Par exemple, si toutes les variables auxiliaires étaient incluses, sauf x_{4i}, la valeur espérée selon le modèle pour les personnes de 17 ans serait la même que celle pour les personnes de 16 ans, au lieu d'être égale à la constante. D'encore meilleurs résultats pourraient être obtenus en utilisant des paramétrisations plus perfectionnées, comme des modèles spline, ce que nous explorerons dans une étude à venir.

Tableau 5
Variables auxiliaires possibles dans la simulation de l'enquête sur la population active

Variable	Définition
x_{1i}	1 (correspond à la constante dans le modèle pour Y)
x_{2i}	1 si la personne i est un homme, -1 si elle est une femme
x_{3i}	1 si la personne i a 16 ans ou plus
x_{4i}	1 si la personne i a 17 ans ou plus
⋮	⋮
$x_{12,i}$	1 si la personne i a 25 ans ou plus
$x_{13,i}$	1 si la personne i a 30 ans ou plus
⋮	⋮
$x_{19,i}$	1 si la personne i a 60 ans ou plus
$x_{20,i}$	x_{3i} si la personne i est un homme $-x_{3i}$ si elle est une femme
⋮	⋮
$x_{36,i}$	$x_{19,i}$ si la personne i est un homme $-x_{19,i}$ si elle est une femme

Nous avons utilisé la sélection séquentielle pour choisir les variables, en partant du modèle ne contenant que le terme constant. À chaque étape, des variables pouvaient être ajoutées ou supprimées, en choisissant celles donnant la meilleure réduction du critère. Quand la sélection séquentielle commençait à tourner en rond (par exemple, ajout de x1, puis ajout de x2, puis suppression de x1, puis suppressions de x2, puis ajout de x1, *etc.*), le processus de construction du modèle était arrêté et le modèle résultant était utilisé comme modèle final. Les estimateurs de Δ pris en considération étaient l'estimateur non robuste, l'estimateur robuste (à l'hétéroscédasticité) et l'estimateur de la variance des grappes finales (UCV) qui est robuste à l'hétéroscédasticité et aux corrélations dans les UPE. Nous n'avons pas utilisé les tests de signification, car ils auraient nécessité l'intégration des corrélations dans les UPE pour être réalistes. Nous ne présentons pas les résultats pour l'estimateur ridge, parce que des points négatifs ont

rarement été produits par cette simulation, de sorte que les résultats étaient fort semblables à ceux des prédicteurs BLUP utilisant toutes les variables auxiliaires.

Résultats

Le tableau 6 donne la RRMSE des divers prédicteurs BLUP adaptatifs et non adaptatifs. L'écart entre les RRMSE du prédicteur BLUP ne contenant que le terme constant et du prédicteur BLUP contenant toutes les variables auxiliaires était assez faible. Il n'est par conséquent pas étonnant que les meilleurs gains mineurs aient été réalisés en utilisant les prédicteurs BLUP adaptatifs plutôt que le prédicteur BLUP contenant toutes les variables. Le prédicteur BLUP adaptatif utilisant l'estimateur $\hat{\Delta}$ non robuste est celui qui a donné la RRMSE la plus faible dans tous les cas.

Tableau 6
RRMSE des prédicteurs de l'emploi dans l'enquête sur la population active

n	BLUP		BLUP adaptatif		
	Toutes	C^{ste}	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	$\hat{\Delta}$ UCV
200	6,54	6,77	6,44	7,06	6,96
400	4,72	4,76	4,61	4,72	4,65
1 000	2,45	2,70	2,43	2,45	2,49

Le tableau 7 donne le nombre moyen de variables sélectionnées pour chacun des prédicteurs BLUP adaptatifs. Des 36 variables auxiliaires possibles, de cinq à neuf environ ont été sélectionnées en se basant sur l'estimateur $\hat{\Delta}$ non robuste. Le nombre de variables sélectionnées augmente parallèlement à la taille d'échantillon. Le critère de robustesse à l'hétéroscédasticité a produit de plus grands ensembles de variables auxiliaires et le critère de variance des grappes finales (UCV) a donné d'encore plus grands ensembles.

Tableau 7
Moyenne (intervalle interquartile) du nombre de variables auxiliaires sélectionnées dans la simulation de la population active

n	Méthode de sélection des variables		
	Non robuste	Robuste	UCV
200	6,5 (5,0- 8,0)	13,4 (10,0-16,0)	16,1 (13,0-19,0)
400	7,4 (6,0- 8,0)	12,1 (9,0-15,0)	14,5 (12,0-17,0)
1 000	8,6 (7,0-10,0)	11,6 (10,0-13,0)	14,2 (12,0-17,0)

Le tableau 8 donne la non-couverture de l'intervalle de confiance (IC) de divers prédicteurs. Les IC à 90 % ont été définis comme étant l'estimateur +/-1,64 erreur-type où la variance a été estimée à l'aide de la méthode de la variance des grappes finales (UVC). L'examen du tableau révèle que le prédicteur BLUP contenant toutes les variables auxiliaires donne lieu à une non-couverture importante pour $n = 200$ et 400. Le prédicteur BLUP adaptatif utilisant l'estimateur $\hat{\Delta}$ non robuste produit une couverture raisonnablement proche de la couverture nominale, tandis que les autres prédicteurs BLUP adaptatifs donnent une non-couverture importante.

Tableau 8
Non-couverture de l'intervalle de confiance (%) des prédicteurs de l'emploi

n	BLUP		BLUP adaptatif		
	Toutes	C^{ste}	$\hat{\Delta}$ non robuste	$\hat{\Delta}$ robuste	$\hat{\Delta}$ UCV
200	17,6	12,0	12,0	20,0	24,0
400	17,2	12,0	14,8	16,8	17,6
1 000	6,4	11,6	7,6	6,8	9,6

Le tableau 9 montre dans quelle mesure les divers poids ont donné de bons résultats quand ils ont été utilisés pour estimer une variable différente, à savoir le chômage (égale à 1 pour les chômeurs et 0 autrement). Nous avons calculé les prédicteurs BLUP adaptatifs en utilisant l'estimateur $\hat{\Delta}$ non robuste et en prenant comme variable d'intérêt l'emploi, ainsi que le chômage. Le prédicteur BLUP adaptatif contenant des variables choisies pour l'emploi avait une RRMSE comprise entre celle du prédicteur BLUP non adaptatif contenant toutes les variables et celle du prédicteur BLUP non adaptatif ne contenant que la constante. Cela donne à penser que ce prédicteur BLUP adaptatif donne des résultats raisonnables, même s'il est appliqué à d'autres variables que l'emploi. En fait, le prédicteur BLUP adaptatif basé sur le chômage avait une RRMSE plus élevée, peut-être parce que les variables auxiliaires n'avaient que peu de pouvoir prédictif, voire aucun, pour le chômage, de sorte qu'essayer d'adapter le choix des variables auxiliaires à cette variable d'intérêt n'a pas donné de bons résultats.

Tableau 9
RRMSE des prédicteurs du chômage dans l'enquête sur la population active

n	BLUP		BLUP adaptatif	
	Toutes	C^{ste}	Basé sur l'emploi	Basé sur le chômage
200	36,3	32,6	34,5	36,0
400	24,1	21,7	22,8	23,7
1 000	14,5	14,2	14,1	14,2

5. Discussion

Les études par simulation décrites ici montrent que les prédicteurs BLUP adaptatifs peuvent donner des améliorations utiles comparativement à de simples options non adaptatives. En ce qui concerne tant les simulations de l'enquête sur les fermes que celles de l'enquête sur la population active, les prédicteurs BLUP adaptatifs basés sur

un estimateur non robuste de Δ et ceux basés sur un test de signification ont les uns et les autres produit une EQMP plus faible que les estimateurs non adaptatifs dans la plupart des cas. Pour ce qui est de l'enquête sur les fermes, les améliorations étaient parfois considérables comparativement à l'option consistant à systématiquement utiliser le modèle complet ou celle consistant à systématiquement utiliser le modèle contenant la constante plus la variable de taille. Dans le cas de l'enquête sur la population active, les améliorations étaient mineures. Les prédicteurs BLUP adaptatifs ont également donné une couverture raisonnable des intervalles de confiance.

Les prédicteurs BLUP adaptatifs basés sur les critères d'estimateur robuste et par la méthode de la variance des grappes finales (UCV) ont donné de nettement moins bons résultats que les autres. Cette constatation est étonnante, car il est connu que les données de l'AAGIS sont hétéroscédastiques et que celles de l'enquête sur la population active sont en grappes, ce qui laisse entendre que l'estimation de la variance par la méthode UCV aurait dû donner de bons résultats. Une analyse plus approfondie de la simulation de l'enquête sur les fermes a montré que, dans la grande majorité des cas, $\hat{\Delta}_{\text{robust}}$ avait une variance plus élevée que $\hat{\Delta}_{\text{nonrobust}}$, particulièrement pour les variables auxiliaires dont le pouvoir prédicteur était faible – voir l'annexe de Clark et Chambers 2008 pour des renseignements plus détaillés. Il semble donc que la méthode robuste ait tendance à aboutir plus fréquemment à la sélection de variables auxiliaires contre-productives, ce qui pourrait expliquer sa performance médiocre.

Nous avons constaté une tendance générale de toutes les méthodes adaptatives à choisir un trop grand nombre de variables auxiliaires, mais malgré cela, les estimateurs adaptatifs ont généralement donné des résultats supérieurs ou comparables à ceux des estimateurs simples non adaptatifs. Selon nous, en pratique, il faudrait conjuguer une recherche automatique de modèle (en utilisant un estimateur $\hat{\Delta}$ non robuste ou un critère de signification statistique) et une certaine forme de jugement subjectif. Par exemple, des modèles pourraient être sélectionnés parmi plusieurs ensembles de variables auxiliaires possibles de tailles différentes. Si les grands ensembles ne donnent que des améliorations apparemment faibles, le statisticien pourrait décider de se limiter à un ensemble plus petit, même s'il paraît légèrement sous-optimal.

Les estimateurs ridge ont aussi donné d'assez bons résultats en ce qui concerne la RRMSE et la couverture de l'intervalle de confiance. Dans l'ensemble, ils ont produit des résultats comparables à ceux des prédicteurs BLUP adaptatifs pour l'estimation du total de la variable d'intérêt quand le choix des variables auxiliaires était fondé sur cette dernière. Cependant, quand nous avons appliqué les poids BLUP adaptatifs à des variables différentes, les estimateurs ridge ont donné des résultats légèrement supérieurs. Une approche encore meilleure pourrait consister à choisir adaptativement les variables auxiliaires qu'il convient d'inclure dans le modèle, ainsi que la façon d'appliquer la régression ridge, en se fondant sur un critère calculé d'après l'échantillon. Cette approche sera le sujet de travaux de recherche à venir.

L'une des réserves qui a été formulée concernant l'approche par prédiction de l'échantillonnage en population finie tient à sa non-robustesse à l'omission de variables auxiliaires importantes. Dans nos simulations s'appuyant sur des données économiques sur les fermes et sur des données sociales, les prédicteurs adaptatifs avaient un biais faible et une erreur quadratique moyenne plus faible que les estimateurs non adaptatifs dans la plupart des nombreux cas examinés dans notre étude par simulation et n'étaient jamais considérablement plus mauvais. À condition que toutes les variables de plan soient considérées comme des variables auxiliaires possibles, le calage adaptatif offre une stratégie robuste et efficace de prédiction en population finie.

Remerciements

Les travaux ont été financés conjointement par l'Australian Research Council et l'Australian Bureau of Statistics. Un rédacteur associé et deux examinateurs ont formulé des commentaires détaillés et judicieux qui nous ont permis d'améliorer considérablement l'article.

Bibliographie

Bardsley, P., et Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33(3), 290-299.

Brewer, K.R.W. (1963). Ratio estimation and finite populations: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Chambers, R., Skinner, C. et Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2), 321-324.

Clark, R.G., et Chambers, R.L. (2008). *Adaptive calibration for prediction of finite population totals*. University of Wollongong Centre for Statistical and Survey Methodology. Consultable à http://www.cssm.uow.edu.au.

Clark, R.G., et Steel, D.G. (2002). The effect of using household as a sampling unit. *Revue Internationale de Statistique*, 70(2), 289-314.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Hansen, M., Madow, W. et Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71(355), 657-664.

Royall, R.M., et Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73(362), 351-358.

Royall, R.M., et Cumberland, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76(373), 66-80.

Royall, R.M., et Cumberland, W.G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76(376), 924-930.

Royall, R.M., et Herson, J. (1973a). Robust estimation in finite populations 1. *Journal of the American Statistical Association*, 68(344), 880-889.

Royall, R.M., et Herson, J. (1973b). Robust estimation in finite populations 2: Stratification on a size variable. *Journal of the American Statistical Association*, 68(344), 890-893.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York : Springer-Verlag.

Silva, P.L.N., et Skinner, C. (1997). Sélection des variables pour l'estimation par régression dans le cas des populations finies. *Techniques d'enquête*, 23, 25-35.

Skinner, C., et Silva, P.L.N. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 76-81.

Smith, T.M.F. (1976). The foundations of survey sampling: A review. *Journal of the Royal Statistical Society*, Series A, 139(2), 183-202.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: A Prediction Approach. New York : John Wiley & Sons, Inc.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York : Springer-Verlag.

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 193-201
Statistique Canada, N° 12-001-X au catalogue

Estimation de la précision d'évolutions dans les enquêtes répétées, application à l'enquête suisse sur la valeur ajoutée

Lionel Qualité et Yves Tillé [1]

Résumé

Nous proposons une méthode pour estimer la variance des estimateurs des évolutions qui prend en compte toutes les composantes de ceux-ci : le plan de sondage, le traitement des non-réponses, le traitement des grosses entreprises, la corrélation de la non-réponse d'une vague à l'autre, l'effet dû à l'utilisation d'un panel, la robustification et le calage au moyen d'un estimateur par le ratio. Cette méthode, qui permet la détermination d'intervalles de confiance des évolutions, est ensuite appliquée à l'enquête suisse sur la valeur ajoutée.

Mots clés : Covariance ; sondage stratifié ; panel.

1. Introduction

Dans les enquêtes longitudinales, la précision des évolutions dépend directement du taux de recouvrement des échantillons. Nous commençons par rappeler des résultats connus concernant les plans simples disjoints (voir à ce sujet Kish 1965 ; Sen 1973 ; Wolter 1985 ; Laniel 1988 ; Hidiroglou, Särndal et Binder (1995) ; Holmes et Skinner 2000 ; Nordberg 2000 ; Fuller et Rao 2001 ; Berger 2004). Ensuite nous calculons la variance des évolutions pour des plans simples dont les échantillons se superposent. Lorsque les taux de sondage sont très faibles la plupart de ces résultats sont bien connus et présentés par exemple dans Caron et Ravalet (2000). On peut trouver des résultats tenant compte des corrections de population finies dans Tam (1984).

Nous avons calculé précisément les variances des estimateurs pour une classe plus large de plans de sondage en population finie. Les corrections de population finie peuvent jouer un rôle important dans les enquêtes auprès des entreprises, car les entreprises de grandes tailles sont parfois sélectionnées avec des probabilités d'inclusion très élevées. Les calculs deviennent beaucoup plus compliqués en population finie pour la raison suivante : si la taille de la population est finie, deux échantillons disjoints ne sont pas indépendants. Si la population est infinie, deux échantillons indépendants sont disjoints. Plusieurs estimateurs sont examinés : la différence des estimateurs transversaux, la différence estimée uniquement sur la partie commune, les évolutions relatives. Les calculs deviennent encore plus complexes lorsque la population est dynamique (naissances, morts, changement de structure). La théorie que nous développons ci-dessous se limite au cas où la population ne change pas au cours du temps.

Dans la première partie, nous présentons le plan de sondage aléatoire simple bidimensionnel (voir à ce sujet Goga 2003) et nous donnons les estimateurs de Horvitz-Thompson correspondants. Nous calculons la variance de l'estimateur des évolutions basé sur ce plan de sondage. Dans une deuxième partie, nous donnons la variance d'autres estimateurs simples : l'évolution relative ou le quotient des totaux, l'estimateur de la différence basé sur l'intersection des échantillons. Nous décrivons ensuite comment ces résultats s'adaptent à la présence de non-réponse ignorable et à l'utilisation d'estimateurs plus complexes, qui font intervenir des poids modifiés pour obtenir des estimateurs calés, ou des variables modifiées par une procédure de robustification.

Ces résultats sur les plans simples sont facilement généralisables aux plans stratifiés à condition que les entreprises ne changent pas de strate d'une vague à l'autre. Enfin, nous appliquons cette méthode à l'enquête suisse sur la valeur ajoutée en prenant en compte toutes les composantes de l'enquête : la stratification, l'effet panel, la non-réponse, la corrélation entre les non-réponses d'une vague à l'autre, le calage au moyen d'un estimateur par ratio, et la robustification.

2. Estimation de la différence dans les plans simples

Soit une population $U = \{1, \ldots, k, \ldots, N\}$ de taille N dans laquelle sont prélevés deux échantillons s_1 et s_2 de tailles respectives n_1 et n_2. Ces échantillons peuvent avoir une partie commune (voir Figure 1).

On suppose que s_1 et s_2 sont des échantillons prélevés selon un plan simple sans remise, les tailles n_1 et n_2 ne sont donc pas aléatoires. Les échantillons s_1 et s_2 peuvent être décomposés en trois parties $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1,$ et $s_C = s_1 \cap s_2$. Soit $n_A = |s_A|, n_B = |s_B|, n_C = |s_C|, n_1 = n_A + n_C, n_2 = n_B + n_C$. Les tailles de $s_A, s_B,$ et $s_C,$ peuvent être

1. Lionel Qualité et Yves Tillé, Institut de Statistique, Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

aléatoires. Ce plan généralise, entre autres, les cas de figure suivants :

- si les échantillons s_1 et s_2 sont sélectionnés indépendamment, n_C est alors une variable aléatoire ;
- si l'échantillon s_1 est d'abord sélectionné, et l'échantillon s_2 est sélectionné dans le complémentaire de s_1 dans U alors s_C est vide et $n_C = 0$;
- si l'échantillon s_1 est d'abord sélectionné, et l'échantillon s_2 est constitué de l'union d'un sous-échantillon de taille fixe de s_1 et d'un échantillon de taille fixe du complémentaire de s_1 dans U, alors n_C n'est pas aléatoire, et l'on se retrouve dans le cas A de Tam (1984).

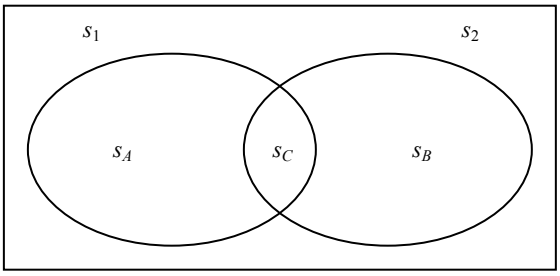


Figure 1 Échantillons qui se chevauchent

On fait l'hypothèse supplémentaire que conditionnellement à $n_A, n_B,$ et $n_C,$ les échantillons $s_A, s_B,$ et $s_C,$ sont simples, sans remise et de taille fixe. Ils proviennent du plan de sondage :

Définition 1. Plan simple bidimensionnel de taille fixe (n_A, n_B, n_C) :

$$p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) = \begin{cases} \dfrac{n_A!\, n_B!\, n_C!\,(N - n_A - n_B - n_C)!}{N!} & \text{si } n_A = |s_A|, \\ & n_B = |s_B|, n_C = |s_C| \\ 0 & \text{sinon,} \end{cases}$$

où $s_A = s_1 \setminus s_2, s_B = s_2 \setminus s_1$ et $s_C = s_1 \cap s_2$ (voir à ce sujet Goga 2003).

La loi de tirage du couple $(s_1, s_2),$ que l'on ne connaît pas en général, est donc supposée être de la forme

$$p(s_1, s_2) = p_{\text{simple}}(s_1, s_2 | n_A, n_B, n_C) \Pr(|s_1 \cap s_2| = n_C).$$

Soit deux variables x et y dont les valeurs prises sur les unités de U sont notées respectivement x_k et $y_k, k \in U$. Les variables x et y peuvent représenter la même variable mesurée à deux moments différents. On suppose également que x ne peut être observée que pour s_1 et y pour s_2. L'objectif est d'estimer les totaux

$$X = \sum_{k \in U} x_k \quad \text{et} \quad Y = \sum_{k \in U} y_k,$$

ainsi que la différence $Y - X$. Les estimateurs de Horvitz-Thompson de X et Y sont donnés par

$$\hat{X}_1 = \frac{N}{n_1} \sum_{k \in s_1} x_k \quad \text{et} \quad \hat{Y}_2 = \frac{N}{n_2} \sum_{k \in s_2} y_k.$$

2.1 Estimation naturelle de la différence

2.1.1 Variance de l'estimation de la différence

Une première manière de procéder pour estimer l'évolution $\Delta = Y - X$ est d'utiliser la différence des estimateurs transversaux $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1$ qui est un estimateur sans biais conditionnellement à n_C sous le plan simple :

$$E(\hat{\Delta} | n_C) = Y - X,$$

et donc est également sans biais sous le plan p non-conditionnel à n_C.

Proposition 1 : La variance de $\hat{\Delta}$ vaut :

$$\begin{aligned} \operatorname{var}(\hat{\Delta}) = {} & N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 \\ & - 2N^2 \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N} \right) S_{xy}, \end{aligned} \tag{1}$$

où

$$S_x^2 = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})^2, S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{Y})^2,$$

$$S_{xy} = \frac{1}{N-1} \sum_{k \in U} (x_k - \bar{X})(y_k - \bar{Y}).$$

La démonstration de (1) se trouve en annexe.

2.1.2 Cas particuliers et gain de précision

Le résultat (1) permet de traiter directement les cas particuliers de coordination suivants :

- si les deux échantillons forment un panel, $n_C = n_1 = n_2,$ alors

  $$\operatorname{var}(\hat{\Delta}) = N^2 \left(\frac{1}{n_C} - \frac{1}{N} \right) (S_x^2 + S_y^2 - 2S_{xy});$$

- si les échantillons sont disjoints (voir aussi Ardilly et Tillé 2003, pages 24-28) $n_C = 0,$ et

  $$\begin{aligned} \operatorname{var}(\hat{\Delta}) = {} & N^2 \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2 + N^2 \left(\frac{1}{n_2} - \frac{1}{N} \right) S_y^2 \\ & + 2NS_{xy}. \end{aligned}$$

  Il est surprenant de constater que la covariance ne dépend pas des tailles des échantillons. Elle est négative si x et y sont positivement corrélées, et devient négligeable par rapport aux termes de

variance quand la taille de la population est grande ;

- si q représente le taux de recouvrement fixé des deux échantillons et que $n_1 = n_2 = n,$ on retrouve le cas A développé par Tam (1984). On obtient alors $n_C = qn,$ et

$$\text{var}(\hat{\Delta}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(S_x^2+S_y^2)-2N^2\left(\frac{q}{n}-\frac{1}{N}\right)S_{xy};$$

- si les deux échantillons sont indépendants, $E(n_C) = n_1 n_2 / N,$ et on retrouve

$$\text{var}_{\text{IND}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)S_x^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)S_y^2.$$

Si la taille de la population est grande et si les variables x et y ont des dispersions proches l'une de l'autre, le gain (ou la perte) de précision dû à la coordination par rapport à la sélection de deux échantillons de manière indépendante est

$$G = \frac{\text{var}(\hat{\Delta})}{\text{var}_{\text{IND}}(\hat{\Delta})} \approx 1-\rho q, \qquad (2)$$

où ρ est le coefficient de corrélation entre x et $y, \rho = S_{xy}/S_x S_y$ et q est le taux de recouvrement, $q = 2E(n_C)/(n_1+n_2).$ L'expression (2) fournit un simple coefficient multiplicatif permettant de prendre en compte l'effet de la corrélation et du recouvrement.

2.1.3 Estimation de la variance de $\hat{\Delta}$

Pour estimer la variance, il faut considérer deux cas :

- si $E(n_C)$ est connu, ce qui peut être le cas (par exemple quand on sait que les deux échantillons sont indépendants), alors

$$\widehat{\text{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{E(n_C)}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}, \qquad (3)$$

où

$$s_{x1}^2 = \frac{1}{n_1-1}\sum_{s_1}(x_k-\bar{x}_1)^2,\ s_{y2}^2 = \frac{1}{n_2-1}\sum_{s_2}(y_k-\bar{y}_2)^2,$$

et

$$s_{xyC} = \frac{1}{n_C-1}\sum_{s_C}(x_k-\bar{x}_C)(y_k-\bar{y}_C).$$

Cet estimateur est sans biais, mais il peut parfois prendre des valeurs negatives ;

- si $E(n_C)$ n'est pas connu, la seule information concernant la coordination est n_C.

$$\widehat{\text{var}}(\hat{\Delta}) = N^2\left(\frac{1}{n_1}-\frac{1}{N}\right)s_{x1}^2 + N^2\left(\frac{1}{n_2}-\frac{1}{N}\right)s_{y2}^2 - 2N^2\left(\frac{n_C}{n_1 n_2}-\frac{1}{N}\right)s_{xyC}. \qquad (4)$$

Cet estimateur est sans biais conditionnellement à n_C et est donc aussi non-conditionnellement sans biais. Il peut aussi prendre parfois des valeurs négatives. Nous verrons plus loin que dans certaines applications où intervient de la non-réponse $E(n_C)$ n'est pas connu.

Pour utiliser l'estimateur (3), il est nécessaire d'avoir au moins deux unités dans l'intersection des échantillons $(n_C \geq 2),$ sauf si $E(n_C) = n_1 n_2 / N.$ En effet, si $E(n_C) = n_1 n_2 / N,$ ce qui est le cas quand les deux échantillons sont indépendants, le troisième terme de l'estimateur (3) est nul. L'estimateur (4) n'est quant à lui pas défini lorsque $n_C = 1,$ sauf si $n_1 n_2 = N.$

2.2 Estimation au moyen de la partie commune

On peut aussi estimer la différence en utilisant uniquement la partie commune de l'échantillon, ce qui donne l'estimateur

$$\hat{\Delta}_C = N(\bar{y}_C - \bar{x}_C),$$

avec $\bar{y}_C = 1/n_C \sum_{k\in s_C} y_k$ et $\bar{x}_C = 1/n_C \sum_{k\in s_C} x_k.$ Cet estimateur est sans biais non-conditionnellement et conditionnellement à n_C.

2.2.1 Estimation de la variance de $\hat{\Delta}_C$

La variance conditionnelle de $\hat{\Delta}_C$ vaut

$$\text{var}(\hat{\Delta}_C | n_C) = N^2\left(\frac{1}{n_C}-\frac{1}{N}\right)(S_y^2 + S_x^2 - 2S_{xy}).$$

La variance non-conditionnelle vaut

$$\text{var}(\hat{\Delta}_C) = N^2\left[E\left(\frac{1}{n_C}\right)-\frac{1}{N}\right](S_y^2 + S_x^2 - 2S_{xy}).$$

Cette variance non-conditionnelle peut être délicate à calculer quand n_C est aléatoire.

2.2.2 Comparaison des variances de $\hat{\Delta}$ et $\hat{\Delta}_C$

Si l'on veut comparer les deux estimateurs de la différence, on peut calculer

$$\text{var}(\hat{\Delta}) - \text{var}(\hat{\Delta}_C) = N^2\left[\frac{1}{n_1} - E\left(\frac{1}{n_C}\right)\right]S_y^2 + N^2\left[\frac{1}{n_2} - E\left(\frac{1}{n_C}\right)\right]S_x^2 - 2N^2\left[\frac{E(n_C)}{n_1 n_2} - E\left(\frac{1}{n_C}\right)\right]S_{xy}.$$

Si $n_1 = n_2 = n, S_x^2 = S_y^2 = S^2,$ et $E(1/n_C) \approx 1/E(n_C),$ alors on obtient

$$\begin{aligned}\mathrm{var}(\hat{\Delta}) - \mathrm{var}(\hat{\Delta}_C) &\approx \frac{1}{qn}[q-1]\,2N^2S^2 - 2\frac{1}{qn}[q^2-1]\,\rho N^2S^2 \\ &= \frac{2N^2S^2}{qn}(1-q)\,[\rho(1+q)-1],\end{aligned}$$

où $q = 2E(n_C)/(n_1+n_2)$ est le taux de recouvrement. L'estimateur $\hat{\Delta}_C$ est donc plus précis que $\hat{\Delta}$ si

$$\rho \geq \frac{1}{1+q}.$$

Par exemple, si $q = 0,7,$ il est préférable de n'utiliser que la partie commune dès que $\rho \geq 1/(1+0,7) \approx 0,588$ (voir à ce sujet Caron et Ravalet 2000, page 346). Dans les cas où le recouvrement est important et la corrélation élevée, l'estimateur basé sur la différence des estimateurs transversaux n'est donc pas très pertinent.

3. Prise en compte de la non-réponse totale

On considère que la non-réponse est indépendante du plan de sélection. Selon le modèle, chaque unité décide de répondre ou non aléatoirement et les probabilités de réponse sont égales entre les unités. Il s'agit du modèle le plus élémentaire. Cependant, si une unité ne répond pas à la première vague, il est fortement probable qu'elle ne répondra pas non plus à la deuxième vague. Le modèle prend en compte cette dépendance en considérant séparément quatre cas :

- l'unité répond à la première vague et à la seconde ;
- l'unité répond à la première vague mais pas à la seconde ;
- l'unité ne répond pas à la première vague mais bien à la seconde ;
- l'unité ne répond ni à la première vague, ni à la seconde.

La non-réponse est couramment modélisée par un plan bernoullien multivarié, ce qui signifie que la probabilité de répondre est la même pour toutes les unités statistiques et également qu'une unité décide de répondre indépendamment de la réponse des autres unités. Le plan de non-réponse est le suivant :

$$q(r_A, r_B, r_C, r_D) = \phi_A^{\mathrm{card} r_A}\phi_B^{\mathrm{card} r_B}\phi_C^{\mathrm{card} r_C}\phi_D^{\mathrm{card} r_D},$$

où $r_A, r_B, r_C, r_D \subset U,$ et r_A, r_B, r_C, r_D sont mutuellement exclusifs, et où

- $\phi_A^{\mathrm{card} r_A}$ est la probabilité de répondre à la vague 1, mais pas à la vague 2 ;
- $\phi_B^{\mathrm{card} r_B}$ est la probabilité de répondre à la vague 2, mais pas à la vague 1 ;
- $\phi_C^{\mathrm{card} r_C}$ est la probabilité de répondre à la vague 1, et à la vague 2 ;
- $\phi_D^{\mathrm{card} r_D}$ est la probabilité de ne répondre ni à la vague 1, ni à la vague 2.

La phase de non-réponse modélisée ainsi consiste donc en la sélection de quatre échantillons disjoints selon des plans bernoulliens avec des intensités différentes. Comme elle est supposée indépendante du plan de sondage, conditionnellement aux tailles d'échantillons observées, le plan résultant de la sélection et de la non-réponse est un plan simple multivarié. Si l'inférence est menée conditionnellement aux tailles d'échantillon, l'estimation des probabilités $\phi_A, \phi_B, \phi_C, \phi_D$ n'est pas nécessaire et une inférence sans biais peut être menée, comme si l'on avait affaire à un plan simple. La théorie de la section précédente s'applique donc directement sur les répondants, et toute l'information sur le recouvrement des deux échantillons se trouve dans $|s_C|,$ que ce recouvrement soit dû au plan ou au lien existant entre les non-réponses aux deux vagues. Remarquons que même si le modèle est assez simple, il prend en compte le fait que si une unité n'a pas répondu à une vague, elle aura probablement moins de chance de répondre à la vague suivante. De plus, ce modèle sera appliqué dans des strates homogènes relativement petites.

4. Autres mesures des évolutions

La mesure de l'évolution n'est pas toujours exprimée en terme de différences. L'évolution est souvent mesurée sous forme de quotient, ou de différence relative. On considère donc les trois mesures suivantes :

- la différence $\hat{\Delta} = \hat{Y}_2 - \hat{X}_1$;
- l'évolution relative $\hat{\Delta}_R = (\hat{Y}_2 - \hat{X}_1)/\hat{X}_1 = \hat{Y}_2/\hat{X}_1 - 1$;
- le quotient $\hat{Q} = \hat{Y}_2/\hat{X}_1.$

La variance de $\hat{\Delta}$ peut s'exprimer simplement en fonction des estimateurs de variance de $\hat{Y}_2$ et $\hat{X}_1$ et de l'estimateur de leur covariance (voir expression 4). La variance de $\hat{\Delta}_R$ est égale à la variance de $\hat{Q}.$ Elles peuvent être approchées puis estimées par une technique de résidus (voir à ce sujet Woodruff 1971 ; Binder et Patak 1994 ; Deville et Särndal 1992 ; Deville 1999),

$$\widehat{\text{var}}(\hat{\Delta}_R) = \widehat{\text{var}}(\hat{Q}) = \frac{1}{\hat{X}_1^2}\left[\widehat{\text{var}}(\hat{Y}_2) + \hat{Q}^2\,\widehat{\text{var}}(\hat{X}_1) - 2\hat{Q}\widehat{\text{cov}}(\hat{X}_1, \hat{Y}_2)\right].$$

Cette variance peut donc être simplement estimée dès lors que l'on dispose d'estimateurs de $\text{var}(\hat{Y}_2)$, $\text{var}(\hat{X}_1)$ et $\text{cov}(\hat{X}_1, \hat{Y}_2)$.

5. Estimation par ratio et robustification

Deux techniques sont couramment utilisées pour les estimations de résultats d'enquêtes par sondage : l'utilisation d'un estimateur par ratio pour caler sur le total d'une variable auxiliaire, et la robustification des estimateurs. Ces techniques doivent être prises en compte pour déterminer la précision des résultats finaux.

5.1 Calage

Si un estimateur est calé sur des totaux connus, la variance peut-être estimée simplement par une technique de résidus (voir Woodruff 1971 ; Binder et Patak 1994 ; Deville et Särndal 1992 ; Deville 1999). Par exemple, si $\mathbf{z}_{k1}$ et $\mathbf{z}_{k2}$ sont deux vecteurs colonnes de variables auxiliaires sur lesquelles les estimateurs $\hat{X}_{1\text{Cal}}$ et $\hat{Y}_{2\text{Cal}}$ sont calés en vagues 1 et 2, alors les variances peuvent être estimées par une technique de résidus : $\text{var}(\hat{X}_{1\text{Cal}}) \approx \text{var}(\hat{E}_1)$ et $\text{var}(\hat{Y}_{2\text{Cal}}) \approx \text{var}(\hat{E}_2)$, où $\hat{E}_1$ et $\hat{E}_2$ sont les estimateurs de Horvitz-Thompson des totaux des résidus, ces derniers étant donnés pour un plan simple et pour l'estimateur par la régression généralisée par :

$$e_{k1} = x_k - \mathbf{z}_{k1}'\hat{\mathbf{B}}_1,$$
$$e_{k2} = y_k - \mathbf{z}_{k2}'\hat{\mathbf{B}}_2,$$

avec

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}\mathbf{z}_{k1}'\right)^{-1}\sum_{k\in s_1} q_{k1}\mathbf{z}_{k1}x_{k1},$$
$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}\mathbf{z}_{k2}'\right)^{-1}\sum_{k\in s_2} q_{k2}\mathbf{z}_{k2}y_{k2},$$

où $q_{kj}, j = 1, 2,$ est un coefficient qui permet de tenir compte d'une éventuelle hétéroscédasticité.

Dans le cas d'un plan de sondage à probabilités inégales, et par exemple d'un plan de sondage stratifié comme l'enquête suisse sur la valeur ajoutée, les résidus s'obtiennent en utilisant une régression pondérée. Il suffit de remplacer $\hat{\mathbf{B}}_1$ et $\hat{\mathbf{B}}_2$ respectivement par

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}\mathbf{z}_{k1}'}{\pi_{k1}}\right)^{-1}\sum_{k\in s_1} \frac{q_{k1}\mathbf{z}_{k1}x_{k1}}{\pi_{k1}}, \quad \text{et} \qquad (5)$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}\mathbf{z}_{k2}'}{\pi_{k2}}\right)^{-1}\sum_{k\in s_2} \frac{q_{k2}\mathbf{z}_{k2}y_{k2}}{\pi_{k2}}, \qquad (6)$$

où π_{kj} est la probabilité d'inclusion de l'unité k dans l'échantillon de la vague j, $j = 1, 2$.

5.2 Robustification

Il est souvent utile d'appliquer une technique de robustification qui permet de traiter les valeurs aberrantes. Considérons simplement que les valeurs aberrantes aient été détectées et que les poids des individus dont les valeurs sont considérées comme aberrantes aient été modifiés par un facteur $u_{kj}(s)$ à la vague j. Ce facteur est compris entre 0 et 1 et est égal à 1 pour les unités qui ont des valeurs considérées comme normales. La variance de l'estimateur robustifié peut être approchée en faisant l'hypothèse classique que les poids $u_{kj}(s)$ ne dépendent que faiblement de l'échantillon s qui a été tiré (voir Hulliger 1999). Il suffit alors de remplacer les variables x_k et y_k observées par $u_{k1}x_k$ et $u_{k2}y_k$ dans les estimateurs de variance.

En remettant ensemble toutes les composantes de l'erreur quadratique moyenne d'une évolution de manière à prendre en compte toutes les composantes de cette variance : le plan, l'effet panel, la non-réponse, le calage et la robustification, on obtient, pour l'évolution relative dans une strate,

$$\widehat{\text{EQM}}(\hat{\Delta}_R) = \widehat{\text{EQM}}(\hat{Q}) = \frac{1}{\hat{X}_1}\left[\widehat{\text{var}}(\widehat{EU}_1) + \hat{Q}^2\,\widehat{\text{var}}(\widehat{EU}_1) - 2\hat{Q}\widehat{\text{cov}}(\widehat{EU}_1, \widehat{EU}_2)\right], \qquad (7)$$

où

$$\hat{X}_1 = \frac{N}{m_1}\sum_{R_1} x_k,\ \ \hat{Y}_2 = \frac{N}{m_2}\sum_{R_2} y_k,\ \ \hat{Q} = \frac{\hat{Y}_2}{\hat{X}_1},$$

$$eu_{k1} = u_{k1}x_k - u_{k1}\mathbf{z}_{k1}'\hat{\mathbf{B}}_1,$$
$$eu_{k2} = u_{k2}y_k - u_{k2}\mathbf{z}_{k2}'\hat{\mathbf{B}}_2,$$

$$\widehat{EU}_j = \frac{N}{m_j}\sum_{R_j} eu_{kj},\ \ \overline{EU}_j = \frac{\widehat{EU}_j}{N},\ \ j = 1, 2,$$

$$\hat{\mathbf{B}}_1 = \left(\sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}\mathbf{z}_{k1}'}{\pi_{k1}}\right)^{-1}\sum_{k\in D_1} \frac{q_{k1}u_{k1}^2\mathbf{z}_{k1}x_k}{\pi_{k1}},$$

$$\hat{\mathbf{B}}_2 = \left(\sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}\mathbf{z}_{k2}'}{\pi_{k2}}\right)^{-1}\sum_{k\in D_2} \frac{q_{k2}u_{k2}^2\mathbf{z}_{k2}y_k}{\pi_{k2}}.$$

$$\widehat{\text{var}}(\widehat{EU}_j) = N^2\left(\frac{1}{m_j}-\frac{1}{N}\right)\frac{1}{m_j-1}\sum_{R_j}(eu_{kj}-\overline{EU}_j)^2,\ j=1,2,$$

$$\widehat{\text{cov}}(\widehat{EU}_1,\widehat{EU}_2) = N^2\left(\frac{m_C}{m_1 m_2}-\frac{1}{N}\right)\frac{1}{m_C-1}\sum_{R_C}(eu_{k1}-\overline{EU}_1) \times (eu_{k2}-\overline{EU}_2).$$

R_1 et R_2 désignent l'ensemble des répondants à la première et à la deuxième vague dans la strate, $m_1 = |R_1|$, $m_2 = |R_2|$, $R_C = R_1 \cap R_2$ et $m_C = |R_1 \cap R_2|$. D_1 et D_2 sont les ensembles de répondants aux deux vagues dans le domaine dans lequel le calage a été réalisé.

6. L'enquête suisse sur la valeur ajoutée

6.1 Présentation de l'enquête

L'enquête suisse sur la valeur ajoutée est une enquête auprès des entreprises, réalisée chaque année. Elle vise à fournir des estimateurs des principaux paramètres de la production en Suisse : la valeur de production brute, le montant des consommations intermédiaires, la valeur ajoutée créée par les entreprises, et le coût de la main d'œuvre. Le plan de sondage utilisé est un échantillonnage stratifié d'entreprises. En 1999, un échantillon de 11 210 exploitations (occupant au moins deux personnes) a été sélectionné et sondé. Cet échantillon a été reconduit en 2000 et en 2001. Il s'agit donc sur cette période d'une enquête par panel. Faute d'un registre d'entreprises permettant d'identifier les naissances et les décès, la population des entreprises a été considérée comme constante pendant cette période. Le seul ajustement sur des données annuelles est réalisé au moyen d'une estimation par le ratio sur le total des Equivalents Temps Plein (ETP) par domaine d'activité, disponible par ailleurs.

La stratification est définie par les deux premiers chiffres de la NOmenclature Générale des Activités économiques (NOGA2) et par la taille de l'entreprise (voir Renfer 2000). Dans chaque strate d'activités, trois strates de taille sont constituées : les petites entreprises employant 2-19 personnes en ETP, les moyennes entreprises, de 20 à M ETP, et les grandes entreprises de plus de M ETP. La strate contenant les grandes entreprises est recensée, tandis que les petites et moyennes entreprises sont sélectionnées aléatoirement avec des taux de sondage différents. La borne M est choisie différemment dans chaque strate d'activités afin d'obtenir une précision optimale. À ces trois vagues, environ 6 000 établissements ont répondu. Le taux de réponse des grandes entreprises, qui devaient être recensées, était proche de 71 % et était plus élevé que celui des petites et moyennes entreprises. Il a été décidé a posteriori de traiter certaines très grosses entreprises séparément selon la méthodologie de la strate 'surprise' de Hidiroglou et Srinath (1981). On peut en effet penser que le taux de réponse a été meilleur pour les entreprises les plus grosses qui ont une structure administrative plus apte à répondre aux questions de l'enquête. Leur appliquer un poids égal à celui des autres grandes entreprises introduirait un biais ainsi qu'une variabilité trop importante. Les poststrates 'surprise' contiennent les 5 % d'exploitations les plus grandes dans le fichier d'enquête. Ces dernières ont alors été considérées comme effectivement recensées et ont reçu le poids 1. Aucun autre traitement (calage, robustification) ne leur sera appliqué. Les strates de tirage de petites, moyennes et grandes entreprises ont été mises à jour et certaines strates, (classes de taille) comptant peu d'exploitations ont été regroupées a posteriori. Si l'on accepte l'hypothèse que les très grandes entreprises ont effectivement été recensées, l'estimateur qui en découle est sans biais, et la variance liée aux très grandes entreprises est nulle. On peut donc calculer uniquement la variance dans les autres strates mises à jour.

Lors de l'enquête, la catégorie d'activités économiques était redemandée aux entreprises. Les estimations sont réalisées au niveau de ces NOGA2 déclarées et non au niveau des NOGA2 de la base de sondage. Un calage sur le nombre d'équivalent temps plein (ETP) donné par le registre des entreprises est ensuite réalisé au moyen d'un estimateur par le quotient au niveau des domaines NOGA2 'déclarés'.

Enfin, une technique de robustification a été utilisée pour écrêter la distribution de certaines variables dans l'échantillon des petites, moyennes et grandes entreprises (voir Hulliger 1999 ; Peters, Renfer et Hulliger 2001). Les poids des établissements dont les valeurs sont considérées comme aberrantes ont été modifiés par un facteur $u_{kj}(s)$ compris entre 0 et 1. Ce facteur est égal à 1 pour les entreprises qui ont des valeurs considérées comme normales.

6.2 Variance de l'évolution de la valeur ajoutée

L'objectif est d'estimer correctement la variance des estimateurs d'évolution de la valeur ajoutée (voir Renfer 2000 ; Peters *et al.* 2001). En calculant les variances sous une hypothèse d'indépendance des échantillons, on surestime largement les variances des évolutions parce que les variables « valeur ajoutée » aux temps t_1 et t_2 sont positivement corrélées. Une prise en compte correcte de tous les aspects du plan de sondage et du redressement devrait fournir de meilleures estimations de la précision. Les travaux portent sur les vagues d'enquête 1999, 2000 et 2001. Entre ces trois dates, l'échantillon brut n'a pas été modifié. Le fait que l'échantillon soit resté fixe devrait permettre d'estimer de manière fiable les évolutions, mais

un taux de réponse proche de 50 % peut faire perdre le bénéfice du panel, pour peu que le nombre de répondants communs aux vagues successives soit faible. Le cas de l'évolution entre deux vagues d'enquête où il y a eu une mise à jour de l'échantillon, et donc deux échantillons bruts et deux populations de référence différents constitue un tout autre problème.

Dans le cas présent, plusieurs raisons contribuent conjointement à l'obtention de faibles variances :

1. *Plan optimal* : Le plan de sondage a été optimisé. Selon la stratification optimale, les grosses entreprises ont des probabilités d'inclusion plus élevées. La strate d'entreprises contribuant le plus à la valeur ajoutée est recensée. Les estimateurs transversaux ont pour cette raison une faible variance.
2. *Fraction de réponse élevée* : Dans la strate recensée des grandes entreprises, le taux de réponse avoisine 70 %. La correction de population finie $(N-n)/N$ peut donc diviser la variance par 3 par rapport au cas d'une population infinie.
3. *Effet panel* : L'échantillon est un panel, ce qui est la meilleure stratégie pour estimer des évolutions.
4. *Corrélation de la non-réponse* : La non-réponse à une vague est très liée à la vague précédente, et ne dégrade donc pas beaucoup le panel.
5. *Corrélation des variables entre les vagues* : Les variables valeurs ajoutées au temps t et $t+1$ sont très corrélées, car il s'agit de la même variable mesurée à deux moments différents.
6. *Calage* : Les estimateurs sont calés dans les strates sur une variable liée à la variable d'intérêt, la variance des estimateurs peut alors s'écrire comme une variance résiduelle.

Sur les 11 210 entreprises sélectionnées en 1999, environ 5 200 ont répondu à la fois en 1999 et en 2000 ; et 5 300 ont répondu aux vagues 2000 et 2001. La taille du panel est donc relativement modeste, et le traitement de la non-réponse va avoir une grande influence sur les résultats. Afin de réaliser des estimations de variance, nous avons fait l'hypothèse que la non-réponse est ignorable (missing completely at random) au sein des strates de tirage.

À chaque vague, les estimations sont réalisées dans les domaines NOGA2 déclarés. Cela implique la possibilité d'un changement de domaine de la part des entreprises, qu'il faudrait essayer de prendre en compte dans les estimations longitudinales. Nous avons décidé de négliger l'impact de ces changements dans un premier temps, et de considérer pour l'estimation de la covariance que les domaines sont fixes et donnés par la valeur déclarée à la première des deux vagues consécutives. Cette simplification n'est pas abusive, dans la mesure où seules 30 entreprises (resp. 25) ont changé de domaine entre 1999 et 2000 (resp. 2000 et 2001), ce qui représente moins de 0,5 % (resp. 0,2 %) des ETP de l'échantillon. Le calage est réalisé pour chaque année, et peut être pris en compte au moyen d'une technique de résidus. Comme pour l'estimation de la précision des estimateurs transversaux, on prend en compte la robustification en repondérant les variables de l'enquête.

Moyennant des hypothèses réalistes, toutes les composantes de la variance peuvent être prises en compte grâce à l'expression générale (7). Cette expression est appliquée au sein de chaque strate et prend en compte toutes les composante de l'enquête sur la valeur ajoutée : l'effet panel, la non-réponse, la stratification, le calage, et la robustification. Les estimateurs de l'enquête sur la valeur ajoutée sont des estimateurs par ratio, et dans ce cas le calcul des résidus est simplifié. En effet, dans le cas du ratio, on calcule les coefficients de régression donnés en (5) et (6) en ayant une seule variable auxiliaire, donc $\mathbf{z}_{kj} = z_{kj}$ est scalaire. De plus, on prend $q_{kj} = 1/z_{kj},$ pour $j = 1, 2,$ et on obtient alors en prenant en compte la robustification :

$$eu_{k1} = u_{k1}x_k - \hat{B}_1 u_{k1} z_{k1},$$
$$eu_{k2} = u_{k2}y_k - \hat{B}_2 u_{k2} z_{k2},$$

où

$$\hat{B}_1 = \frac{\sum_{D_1} u_{k1}x_k / \pi_{k1}}{\sum_{D_1} u_{k1}z_{k1} / \pi_{k1}},$$
$$\hat{B}_2 = \frac{\sum_{D_2} u_{k2}y_k / \pi_{k2}}{\sum_{D_2} u_{k2}z_{k2} / \pi_{k2}}.$$

6.3 Estimation de la précision des évolutions

Nous avons réalisé des estimations des écarts-types des évolutions des valeurs de productions brutes et des valeurs ajoutées calculées par l'Office fédéral suisse de la statistique. Ces estimations tiennent compte de tous les aspects développés précédemment. Nous les avons comparés aux écarts-types estimés qui auraient été obtenus sous l'hypothèse d'indépendance de tirage entre les vagues. Sur l'ensemble des strates d'activités, les écarts-types qui prennent en considération la corrélation entre les vagues d'enquête sont de 41 % inférieurs à ceux calculés sous hypothèse d'indépendance. Cela permet donc d'avoir des intervalles de confiance bien plus petits que ceux calculés avant ce travail, d'une manière plus rapide mais moins précise. Le gain n'est cependant pas le même dans toutes les strates d'activités. Dans les tableaux suivants sont donnés les écarts-types (ET) calculés pour les cinq plus grosses strates d'activités (NOGA) des évolutions de la valeur de

production brute (ΔVP) et de la valeur ajoutée (ΔVA) entre 1999 et 2000. L'écart-type qui aurait été obtenu en négligeant la corrélation entre les échantillons (ET_{ind}) est également inclus dans les tableaux, ainsi que le « gain » de précision réalisé en tenant compte de cette corrélation.

Tableau 1
Évolution de la valeur de production brute entre 1999 et 2000 et écarts-types (en milliards de francs suisses)

Strate	ΔVP	ET_{ind}	ET	Gain (en %)
1	3,31	2,35	0,87	63
2	-0,77	4,38	1,98	55
3	3,07	2,11	0,94	56
4	4,33	1,10	1,00	09
5	-0,09	0,81	0,53	35

Tableau 2
Évolution de la valeur ajoutée entre 1999 et 2000 et écarts-types (en milliards de francs suisses)

Strate	ΔVA	ET_{ind}	ET	Gain (en %)
1	1,96	0,91	0,32	65
2	0,68	2,99	1,04	65
3	1,90	1,47	0,72	51
4	0,36	0,47	0,45	05
5	-0,36	0,59	0,43	27

Remerciements

Ce travail a été réalisé dans le cadre d'une convention entre l'Université de Neuchâtel et l'Office fédéral suisse de la statistique. Les résultats publiés dans cet article n'engagent que les auteurs et en aucun cas l'Office fédéral de la statistique. Nous remercions Paul-André Salamin pour sa contribution à ce travail.

Annexe

Démonstration de la proposition 1

Il est bien connu que

$$\text{var}(\hat{X}_1) = N^2\left(\frac{1}{n_1} - \frac{1}{N}\right)S_x^2$$

et

$$\text{var}(\hat{Y}_2) = N^2\left(\frac{1}{n_2} - \frac{1}{N}\right)S_y^2.$$

Il suffit donc de calculer $\text{cov}(\hat{X}_1, \hat{Y}_2)$. On note

$$\overline{x}_A = \frac{1}{n_A}\sum_{k\in s_A} x_k, \qquad \overline{x}_C = \frac{1}{n_C}\sum_{k\in s_C} x_k,$$

$$\overline{y}_B = \frac{1}{n_B}\sum_{k\in s_B} y_k, \qquad \overline{y}_C = \frac{1}{n_C}\sum_{k\in s_C} y_k,$$

$$\overline{x}_1 = \frac{n_A\overline{x}_A + n_C\overline{x}_C}{n_1}, \quad \overline{y}_2 = \frac{n_B\overline{y}_B + n_C\overline{y}_C}{n_2},$$

alors $\hat{X}_1 = N\overline{x}_1$ et $\hat{Y}_2 = N\overline{y}_2$. Il reste à calculer

$$\text{cov}(\overline{x}_1, \overline{y}_2) = E\,\text{cov}(\overline{x}_1, \overline{y}_2 \mid n_A, n_B, n_C) + \text{cov}\left[E(\overline{x}_1 \mid n_A, n_B, n_C), E(\overline{y}_2 \mid n_A, n_B, n_C)\right].$$

Comme $\overline{x}_1$ et $\overline{y}_2$ sont sans biais conditionnellement à $n_A, n_B,$ et $n_C,$

$$\text{cov}\left[E(\overline{x}_1 | n_A, n_B, n_C), E(\overline{y}_2 | n_A, n_B, n_C)\right] = \text{cov}(\overline{X}, \overline{Y}) = 0.$$

On obtient donc

$$\text{cov}(\overline{x}_1, \overline{y}_2) = E\,\text{cov}(\overline{x}_1, \overline{y}_2 | n_A, n_B, n_C).$$

Conditionnellement à $n_A, n_B,$ et $n_C,$ on est dans le cas A de Tam (1984, théorème 1). La variance conditionnelle vaut :

$$\text{cov}(\overline{x}_1, \overline{y}_2 | n_A, n_B, n_C) = \left(\frac{n_C}{n_1 n_2} - \frac{1}{N}\right)S_{xy}$$

et donc

$$\text{cov}(\overline{x}_1, \overline{y}_2) = \left(\frac{E(n_C)}{n_1 n_2} - \frac{1}{N}\right)S_{xy}.$$

Or

$$\text{cov}(\hat{X}_1, \hat{Y}_2) = N^2\,\text{cov}(\overline{x}_1, \overline{y}_2),$$

ce qui permet d'obtenir le résultat (1).

Bibliographie

Ardilly, P., et Tillé, Y. (2003). *Exercices corrigés de méthodes de sondage*. Paris : Ellipses.

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 4, 451-467.

Binder, D.A., et Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Caron, N., et Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête Emploi en continu. Rapport technique 0005. Méthodologie Statistique, INSEE, Paris.

Deville, J.-C. (1999). Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus. *Techniques d'enquête*, 25, 219-230.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Fuller, W.A., et Rao, J.N.K. (2001). Un estimateur composite de régression qui s'applique à l'Enquête sur la population active du Canada. *Techniques d'enquête*, 27, 49-56.

Goga, C. (2003). Estimation de la variance dans les sondages à plusieurs échantillons et prise en compte de l'information auxiliaire par des modèles nonparamétriques. Thèse de doctorat, Université de Rennes II, Haute Bretagne, France.

Hidiroglou, M., Särndal, C.-E. et Binder, D. (1995). Weighting and Estimation in Business Surveys. *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M. Colledge et P.S. Kott), New York : John Wiley & Sons, Inc., 477-502.

Hidiroglou, M.A., et Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 690-695.

Holmes, D.J., et Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change. Technical report, Government Statistical Service Methodology Series, 21, Londres, Angleterre.

Hulliger, B. (1999). Simple and robust estimators for sampling. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 54-63.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Laniel, N. 1988. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 246-250.

Nordberg, L. (2000). On variance estimation for measure of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée : procédure d'extrapolation des données. Rapport technique, Office fédéral suisse de la statistique.

Renfer, J.-P. (2000). Enquête sur la production et la valeur ajoutée : échantillonnage complémentaire. Rapport technique, Office fédéral suisse de la statistique.

Sen, A.R. (1973). Theory and application of sampling on repeated occasions with several auxiliary variables. *Biometrics*, 29, 381-385.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, (4), 288-289.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Spinger-Verlag.

Woodruff, R.S. (1971). A simple method for approximating de variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 203-216
Statistique Canada, N° 12-001-X au catalogue

Masquage des UPE et estimation de la variance dans les enquêtes complexes

Inho Park [1]

Résumé

L'analyse des données recueillies auprès d'un échantillon stratifié à plusieurs degrés requiert de l'information sur le plan de sondage, telle que les identificateurs de strate et d'unité primaire d'échantillonnage (UPE), ou les poids de rééchantillonnage connexes, pour l'estimation de la variance. Dans certains fichiers de données à grande diffusion, l'information sur le plan de sondage est masquée en vue d'éviter le risque de divulgation, tout en permettant à l'utilisateur d'obtenir des estimations valides des variances. Par exemple, dans le cas des enquêtes aréolaires comptant un nombre limité d'UPE, les UPE originales sont divisées et (ou) recombinées pour construire des pseudo-UPE dans lesquelles sont permutées les unités d'échantillonnage de deuxième degré et de degré subséquent. Cependant, ces méthodes de masquage des UPE faussent manifestement la structure de mise en grappes du plan d'échantillonnage, ce qui donne des estimations de variance biaisées pouvant présenter un rapport systématique entre les deux estimations de variance obtenues avec et sans masquage des identificateurs d'UPE. Certains travaux antérieurs ont révélé certaines tendances du ratio des estimations de la variance obtenues avec et sans masquage si on représente ce ratio graphiquement en fonction de l'effet de plan sans masquage. Le présent article traite de l'effet du masquage des UPE sur les estimations de la variance sous échantillonnage en grappes en fonction de divers aspects, dont la structure de mise en grappes et le degré de masquage. En outre, nous tâchons d'établir une stratégie de masquage des UPE par permutation des unités d'échantillonnage du degré subséquent qui réduit le biais résultant des estimations de la variance. En guise d'illustration, nous utilisons des données provenant de la National Health Interview Survey (NHIS) auxquelles nous avons apporté certaines modifications artificielles. La stratégie proposée permet de bien réduire le biais des estimations de la variance. Les résultats tant théoriques qu'empiriques indiquent que l'effet du masquage des UPE sur les estimations de la variance est modeste si la permutation des unités d'échantillonnage de degré subséquent est minimale. Nous avons appliqué la stratégie de masquage proposée aux données diffusées de la National Health and Nutrition Examination Survey (NHANES) de 2003-2004.

Mots clés : Contrôle de la divulgation ; échantillonnage stratifié à plusieurs degrés ; permutation des unités d'échantillonnage de degré subséquent ; effet de plan ; coefficient de corrélation intragrappe (CCI) ; moyenne d'échantillon.

1. Introduction

L'analyse des données recueillies auprès d'un échantillon stratifié à plusieurs degrés requiert l'utilisation d'information sur le plan, telle que les identificateurs de strate et d'unité primaire d'échantillonnage (UPE), ou les poids de rééchantillonnage connexes, pour l'estimation de la variance. Dans les grandes enquêtes, les UPE correspondent souvent à un ou à plusieurs comtés. Certaines sources externes mises à la disposition du public, telles que les données du recensement, peuvent fournir des renseignements démographiques très détaillés au niveau de l'UPE. Même si l'on supprime leur nom, l'inclusion des identificateurs d'UPE dans les fichiers de données à grande diffusion peut à elle seule poser un risque d'identification en permettant leur couplage à des données de source externe. Par conséquent, les identificateurs d'UPE sont souvent masqués en vue de 1) réduire le risque de divulgation des données et 2) permettre néanmoins à l'utilisateur d'obtenir des estimations de variance valides. Mayda, Mohl et Tambay (1996) abordent la question du risque éventuel de divulgation des données associé à l'inclusion des identificateurs originaux d'UPE dans les fichiers de données à grande diffusion et considèrent la méthode de regroupement de strates proposée par Rust (1986) afin de trouver un juste équilibre entre les deux exigences susmentionnées. Étant donné la non-convergence éventuelle de l'estimation de la variance sous la méthode de regroupement de strates signalée par Valliant (1996), Yung (1997) a proposé de construire un ensemble de poids de rééchantillonnage bootstrap moyens. Lu (2004) a démontré que fournir des poids de rééchantillonnage et donner les identificateurs de strate et d'UPE sont des options à peu près équivalentes en ce qui concerne la confidentialité, puisqu'il est facile d'obtenir les uns à partir des autres. Shah (2001) a examiné divers moyens de créer des pseudo-strates et des pseudo-UPE étant donné un ensemble de poids obtenus par la méthode des répliques répétées équilibrées. Eltinge (1999) a proposé une méthode semblable à celle du regroupement des strates. Lu, Brick et Sitter (2006) ont également établi les conditions de convergence de l'estimateur de la variance sous la méthode de regroupement des strates et ont proposé des algorithmes de regroupement de strates produisant des estimateurs de variance efficaces et convergents.

1. Inho Park, statisticien, Département des statistiques économiques, Banque de la Corée, Namdaemun-Ro 106, Jung-Gu, Séoul 100-794, Corée. Courriel : ipark@bok.or.kr.

Quand le nombre d'UPE dans l'échantillon est limité, la méthode de regroupement des strates n'est pas intéressante à cause du nombre insuffisant de degrés de liberté pour l'estimation de la variance. Dohrmann, Curtin, Mohadjer, Montaquila et Le (2002), ainsi que Dohrmann, Lu, Park, Sitter et Curtin (2005) ont abordé ce genre de situation et considéré deux méthodes de masquage des UPE. La première consiste à diviser chaque UPE en deux pseudo-UPE (ensembles d'unités finales d'échantillonnage dans les UPE), en doublant arbitrairement le nombre de degrés de liberté pour l'estimation de la variance. La deuxième méthode consiste à construire les pseudo-UPE par permutation des unités d'échantillonnage de deuxième degré (USE, unités secondaires d'échantillonnage) entre les UPE originales, en gardant les nombres originaux de degrés de liberté pour l'estimation de la variance. Autrement dit, les affectations aux UPE et aux strates de toutes les unités finales d'échantillonnage comprises dans une USE sont remplacées par celles dans l'USE appariée. Cette méthode peut être généralisée de manière que les UPE originales soient divisées en une ou plusieurs fractions qui sont ensuite recombinées pour construire des pseudo-UPE dans lesquelles les fractions d'UPE sont permutées. Cette approche diffère de la permutation des données (Dalenius et Reiss, 1982), qui est souvent utilisée pour empêcher la divulgation grâce à la permutation des valeurs des variables d'enquête de nature délicate entre enregistrements individuels. À cause de la distorsion résultante de la structure de mise en grappes du plan d'échantillonnage, les deux méthodes de masquage des UPE peuvent donner des estimations de la variance biaisées présentant éventuellement un rapport systématique entre les deux estimations de la variance obtenues avec et sans masquage des identificateurs d'UPE. Dohrmann et coll. (2005) ont observé des courbes en forme d'entonnoir décroissantes pour le ratio des estimations des variances avec et sans masquage d'une moyenne d'échantillon, quand ces estimations de variance sont représentées graphiquement en fonction de l'effet de plan. Ils ont expliqué ces courbes en se basant sur une relation approximative de l'estimation de la variance, monotone en le coefficient de corrélation intragrappe (CCI) de la formule de l'effet de plan de Kish.

Dans le présent article, nous nous penchons sur les questions soulevées par la deuxième méthode de masquage des UPE qui consiste à permuter les unités d'échantillonnage de degré subséquent entre les UPE originales et nous examinons les effets de cette méthode sur les estimations de la variance. À la section 2, nous étudions l'effet du masquage des UPE sur la variance en ce qui concerne certains aspects de la structure de mise en grappes, comme le CCI et les moyennes et les tailles des fractions d'UPE pour la permutation sous un plan d'échantillonnage en grappes à un degré. À la section 3, nous examinons la relation entre le degré de permutation pour le masquage des UPE et le biais dans la variance en utilisant un modèle paramétrique pour l'échantillonnage en grappes. À la section 4, nous envisageons une stratégie de masquage des UPE par permutation des USE qui aide à réduire les effets du masquage des UPE sur les estimations de la variance et, donc, les biais résultants dans le cas des enquêtes complexes. À la section 5, nous examinons brièvement les travaux récents de Dohrmann et coll. (2002, 2005) et nous présentons les résultats de l'application de la stratégie de masquage proposée à des données provenant de la National Health Interview Survey (NHIS) en y apportant certaines modifications artificielles. Enfin, à la section 6, nous présentons une discussion.

2. Effet de la distorsion de la structure de mise en grappes sur la variance de la moyenne d'échantillon

L'échantillonnage en grappes, souvent utilisé dans les enquêtes pour des raisons économiques et logistiques, est une source importante d'accroissement de la variance d'un estimateur comparativement à un échantillonnage aléatoire simple qui tient à la similarité des unités d'échantillonnage dans les grappes. Les traités classiques d'échantillonnage, tels que Särndal, Swensson et Wretman (1992, section 8.7) fournissent des formules pour calculer la variance d'une moyenne d'échantillon en fonction du CCI, des tailles de grappe et des moyennes d'une variable étudiée y. Cela signifie que l'effet de la mise en grappes du plan d'échantillonnage sur la variance devrait être révélé par l'entremise de ces éléments. À la présente section, nous examinons comment la distorsion de la structure de mise en grappes du plan d'échantillonnage influe sur la variance d'une moyenne d'échantillon quand les UPE sont masquées par la permutation de leurs fractions entre deux UPE. Pour les besoins de la discussion présentée dans cette section, nous considérons un échantillonnage des UPE à un degré avec probabilité proportionnelle à la taille (PPT). Ce plan d'échantillonnage est relativement simple, mais néanmoins suffisamment complexe pour révéler l'effet du masquage des UPE sur la variance en ce qui a trait aux trois éléments susmentionnés.

2.1 Variance sous échantillonnage en grappes PPT à un degré

Supposons qu'une population U de M unités est groupée en N UPE de M_i unités chacune. Nous tirons un échantillon aléatoire de n UPE avec les probabilités p_i ($\sum_{i=1}^{N} p_i = 1$) et nous incluons dans l'échantillon toutes les unités comprises dans une UPE échantillonnée. Pour

simplifier, nous supposons que la sélection des UPE est faite avec remise. La moyenne d'échantillon pondérée $\hat{\bar{Y}} = (\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} w_{ij}\, y_{ij}$ est un estimateur de la moyenne de population $\bar{Y} = M^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$ de la variable étudiée y, où $w_{ij} = (np_i)^{-1}$ et y_{ij} désignent le poids d'échantillonnage et la valeur de y pour la j^{e} unité de l'UPE i, respectivement. Soit $m = \sum_{i=1}^{n} M_i$, $S_y^2 = (M-1)^{-1}$ $\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})^2$, $\bar{Y}_i = M_i^{-1}\sum_{j=1}^{M_i} y_{ij}$ la taille d'échantillon, la variance de population et les moyennes de y dans les UPE, respectivement. En supposant que N est grand de sorte que $N/(N-1) \doteq 1$, sa variance approximative peut s'écrire

$$
\begin{aligned}
V(\hat{\bar{Y}} \mid S) &\doteq m^{-1} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\
&\quad + (mN)^{-1}\sum_{i=1}^{N} p_i^{-1} M_i (M_i M^{-1} - p_i)(\bar{Y}_i - \bar{Y}), \\
&\doteq (mMN)^{-1}\sum_{i=1}^{N} p_i^{-1} M_i^2 (\bar{Y}_i - \bar{Y})^2, \qquad (1)
\end{aligned}
$$

où S désigne l'ensemble d'indices dans l'échantillon, $\rho_{yU} = 1 - S_{yw}^2/S_y^2$ est le CCI et $S_{yw}^2 = (M-N)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y}_i)^2$ est l'écart quadratique moyen dans les UPE. La preuve de (1) est donnée en annexe.

Pour un cas particulier commun de $p_i \propto M_i$, c'est-à-dire l'échantillonnage PPT, (1) se simplifie en

$$
\begin{aligned}
V(\hat{\bar{Y}} \mid S) &\doteq m^{-1} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\
&\doteq (mN)^{-1}\sum_{i=1}^{N} M_i (\bar{Y}_i - \bar{Y})^2 \qquad (2)
\end{aligned}
$$

et le CCI est exprimé par

$$
\rho_{yU} \doteq (\bar{M} - 1)^{-1}\left[\frac{V(\hat{\bar{Y}} \mid S) - m^{-1} S_y^2}{m^{-1} S_y^2}\right]. \qquad (3)
$$

Dans (2), la deuxième approximation indique que les UPE pour lesquelles le terme $M_i(\bar{Y}_i - \bar{Y})^2$ est grand contribuent davantage à la variance. Dans (3), le CCI révèle (approximativement) la perte de précision due à l'accroissement relatif par grappe de la variance de $m^{-1}S_y^2$, la variance de la simple moyenne d'échantillon $\bar{y} = m^{-1}\sum_{(ij)\in S} y_{ij}$ qui aurait pu être obtenue à partir des échantillons aléatoires simples avec remise de même taille.

Une enquête complexe comporte souvent l'échantillonnage PPT à un degré susmentionné ou un autre plan complexe (additionnel) (par exemple, stratification, échantillonnage à plusieurs degrés et probabilités de sélection inégales), ou des caractéristiques d'estimation (par exemple, corrections pour la non-réponse et corrections par calage). Par exemple, si p_i n'est pas proportionnelle à la taille ou qu'on a procédé à un sous-échantillonnage poussé qui aurait pour résultat d'obtenir des poids de sondage inégaux, la caractéristique complexe correspondante pourrait joué un rôle dans l'estimation de la variance. L'incidence connexe sur l'estimation de la variance dans le cas d'échantillons complexes sera discutée en détail à la section 4.

2.2 Moyennes et tailles des fractions d'UPE

Afin de masquer les UPE, considérons que les deux premières UPE dans l'échantillon sont divisées chacune en deux ensembles d'unités, $U_1 = U_{11} \cup U_{12}$ et $U_2 = U_{21} \cup U_{22}$, disons, et que les deux pseudo-UPE, $U_1^* = U_{11} \cup U_{22}$ et $U_2^* = U_{21} \cup U_{12}$, sont construites par permutation de U_{12} et U_{22} entre les deux UPE. Soit S et S^* les ensembles d'indices dans l'échantillon sans masquage et avec masquage, respectivement. Soit $V(\hat{\bar{Y}} \mid S^*)$ la variance de $\hat{\bar{Y}}$ associée aux pseudo-UPE (en supposant aussi que les UPE non échantillonnées dans U restent les mêmes). En outre, soit $\bar{Y}_i^*$, Y_i^*, M_i^* la moyenne et le total de y et la taille de la i^{e} pseudo-UPE, respectivement. En supposant que $\bar{Y} = 0$ sans perte de généralité, la différence entre les variances avec et sans masquage s'écrit, en partant de (2), sous la forme

$$
\begin{aligned}
&V(\hat{\bar{Y}} \mid S^*) - V(\hat{\bar{Y}} \mid S) \\
&\doteq (mN)^{-1}\sum_{i=1,2}(M_i^* \bar{Y}_i^{*2} - M_i \bar{Y}_i^2), \\
&\doteq (mN)^{-1}\sum_{i=1,2}\left(\sqrt{M_i^*}\,\bar{Y}_i^* + \sqrt{M_i}\,\bar{Y}_i\right)\left(\sqrt{M_i^*}\,\bar{Y}_i^* - \sqrt{M_i}\,\bar{Y}_i\right). \qquad (4)
\end{aligned}
$$

L'expression (4) montre que la différence de variance due au masquage des UPE dépend des variations des quantités d'UPE $\sqrt{M_i}\bar{Y}_i$. Si Y_{il} et M_{il} désignent le total et la taille de la fraction d'UPE U_{il}, respectivement, pour i, l = 1, 2, alors $Y_i = Y_{i1} + Y_{i2}$, $Y_i^* = Y_{i1} + Y_{i'2}$, $M_i = M_{i1} + M_{i2}$ et $M_i^* = M_{i1} + M_{i'2}$ pour $i \neq i' = 1, 2$. Il découle clairement de (4) que la variance ne variera pas sous masquage des UPE si la condition suivante est vérifiée :

$$
\sqrt{M_i}\bar{Y}_i = \sqrt{M_i^*}\bar{Y}_i^* \quad \text{ou} \quad \frac{Y_{i1} + Y_{i2}}{\sqrt{M_{i1} + M_{i2}}} = \frac{Y_{i1} + Y_{i'2}}{\sqrt{M_{i1} + M_{i'2}}}. \qquad (5)
$$

Ce résultat est un peu surprenant, puisque par intuition naïve, on pourrait penser que les fractions d'UPE pour la permutation avec $\bar{Y}_i = \bar{Y}_i^*$ préserveront la variance. Pour mieux comprendre (5), considérons les trois cas qui suivent. Si $M_{12} = M_{22}$, alors (5) implique que $Y_{12} = Y_{22}$ ou $\bar{Y}_{12} = \bar{Y}_{22}$, où $\bar{Y}_{il} = Y_{il}/M_{il}$ désigne la moyenne de U_{il}. Si $Y_{12} = Y_{22}$, alors (5) implique que $M_{12} = M_{22}$ ou $\bar{Y}_{12} = \bar{Y}_{22}$. Si $\bar{Y}_{12} = \bar{Y}_{22}$, alors (5) peut s'écrire sous la forme $M_{i1}\bar{Y}_{i1}(\sqrt{M_i^*/M_i} - 1) + \bar{Y}_{i2}(M_{i2}\sqrt{M_i^*/M_i} - M_{i'2}) = 0$ pour $i = 1, 2$, expression qui tient quand $M_{12} = M_{22}$. Ces trois cas démontrent clairement que la variance ne varie pas si les fractions d'UPE pour la permutation sont formées de manière que leurs tailles et leurs moyennes soient les unes et les autre égales.

2.3 Variation du CCI

L'effet de la distorsion de la structure de mise en grappes sur la variance peut aussi être étudié à l'aide du ratio des variances avec et sans masquage. Soit ρ^*_{yU} le CCI avec masquage, c'est-à-dire le CCI défini par les identificateurs d'UPE masqués. De (3), il apparaît clairement que la différence entre les CCI avec et sans masquage est proportionnelle à la différence entre les variances correspondantes, c'est-à-dire

$$\rho^*_{yU} - \rho_{yU} \doteq [m^{-1}(\bar{M}-1)S_y^2]^{-1}[V(\hat{\bar{Y}} \mid S^*) - V(\hat{\bar{Y}} \mid S)].$$

Dans (2), la deuxième approximation indique que la variation du CCI dépend de la façon dont les fractions d'UPE sont formées pour la permutation, c'est-à-dire de la variation de $M_i(\bar{Y}_i - \bar{Y})^2$. Partant de (2), le ratio des variances avec et sans masquage est donné par

$$RV(\hat{\bar{Y}} \mid S,\ S^*) \doteq \frac{\rho^*_{yU} + (\bar{M}-1)^{-1}}{\rho_{yU} + (\bar{M}-1)^{-1}}. \qquad (6)$$

Voir, aussi, Dohrmann et coll. (2005, équation 8). Sous la relation $\rho^*_{yU} = c_y \rho_{yU}$ pour tout $c_y > 0$ donné, (6) est monotone en ρ_{yU}. De plus, le ratio est très instable quand ρ_{yU} ou ρ^*_{yU} est proche de $-(\bar{M}-1)^{-1}$, la borne inférieure du CCI, parce que le numérateur et le dénominateur ayant tous deux leur CCI proche de la borne inférieure, ils sont tous deux proches de zéro. Cela signifie que toute variable de ce type sera fortement influencée par le masquage des UPE.

En général, les enquêtes ont pour but de recueillir des données sur plus d'une variable et le masquage des UPE fondé sur une seule variable pourrait donc ne pas bien protéger le CCI et, par conséquent, la variance d'autres variables. Pour mieux comprendre cet aspect, considérons les situations où le masquage des UPE produit une distorsion à la fois fixe et aléatoire du CCI exprimée sous la forme $\rho^*_{yU} = c_y \rho_{yU} + e$ pour $-0,02 < \rho_y < 0,2$, où $c_y = 0,7,\ 1,0,\ 1,3$, $e \sim N(0,\ 0,05^2)$ et $m = \bar{M} = S_y^2 = 100$. Dans le modèle, le coefficient constant c_y et le terme d'erreur e tiennent compte, respectivement, de la perturbation déterministe et de la perturbation aléatoire causées par le masquage dans le CCI de la variable correspondante. La figure 1 donne le ratio résultant des erreurs-types (racines carrées des variances) avec et sans masquage en fonction du CCI du plan d'échantillonnage. Les trois diagrammes de dispersion de la figure 1 ont une forme en entonnoir similaire et révèlent une forte variation pour les valeurs très faibles de ρ_{yU}. Cependant, leurs courbes génériques dépendent de la grandeur de c_y. Par exemple, $c_y < 1$ produit une courbe décroissante, $c_y > 1$, une courbe croissante et $c_y = 1$, une courbe non monotone. Comme nous en discuterons à la section 3.2, le cas de $c_y > 1$ pourrait être très rare.

La discussion qui précède ne s'étend pas facilement à d'autres situations d'enquête complexe, principalement parce que les enquêtes comportent souvent des caractéristiques de plan de sondage complexes, telles que la stratification, l'échantillonnage à trois degrés ou plus et l'échantillonnage avec probabilités inégales. Dans de telles circonstances, il pourrait être difficile de définir le CCI et de bien approximer la variance sous la forme de (2) (voir, par exemple, Park (2004) et les références qu'il mentionne). Néanmoins, la discussion de la présente section contribue à la compréhension de l'effet du masquage des UPE sur les estimations de la variance en général.

3. Effet du degré de masquage des UPE sur la variance de la moyenne d'échantillon

Le biais dans l'estimation de la variance est d'autant plus important que la structure de mise en grappes est faussée. Pour étudier cette relation, nous considérons un modèle paramétrique utilisé pour l'échantillonnage à deux degrés. Supposons que l'échantillonnage à deux degrés consiste à sélectionner n UPE et m unités dans chaque UPE échantillonnée. À l'exemple de Valliant, Dorfman et Royall (2000, page 248), nous supposons qu'une valeur échantillonnée y_{ij} pour la j^{e} unité de l'UPE i est produite à l'aide du modèle suivant :

$$\xi: E_\xi(y_{ij}) = \mu_{yi} \ \&\ \mathrm{Cov}_\xi(y_{ij},\ y_{i'j'}) = \begin{cases} \sigma^2_{yi} & \text{si } i = i',\ j = j', \\ \sigma^2_{yi}\rho_{yi} & \text{si } i = i',\ j \neq j', \\ 0 & \text{autrement,} \end{cases}$$

où μ_{yi}, σ^2_{yi} et ρ_{yi} sont la moyenne, la variance et la corrélation des unités dans l'UPE i, respectivement. La variance d'une moyenne d'échantillon $\bar{y}_{2st} = (nm)^{-1}\sum_{i=1}^n \sum_{j=1}^m y_{ij}$ s'écrit

$$V_\xi(\bar{y}_{2st} \mid S) = (nm)^{-1}\sigma^2_{yu}[1 + (m-1)\rho_{yu}], \qquad (7)$$

où $\sigma^2_{yu} = n^{-1}\sum_{i=1}^n \sigma^2_{yi}$, $\rho_{yu} = \sum_{i=1}^n(\sigma^2_{yi}\rho_{yi})/\sum_{i=1}^n \sigma^2_{yi}$ et S désigne l'ensemble d'indices dans l'échantillon. Notons que ρ_{yu} peut être interprété comme étant le CCI (groupé ou pondéré par σ^2_{yi}) sous le modèle ξ.

Soit β la taille relative des fractions d'UPE qui doivent être permutées entre les UPE i_1 et i_2. Pour simplifier, nous supposons que $m\beta$ est une valeur entière, qui est le nombre d'unités dans chaque fraction utilisée pour la permutation. Soit S^* l'ensemble d'indices dans l'échantillon avec masquage. La variance de $\bar{y}_{2st}$ avec S^* peut s'écrire

$$V_\xi(\bar{y}_{2st} \mid S^*) = V_\xi(\bar{y}_{2st} \mid S) + n^{-2}(\gamma - 1)(\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2}) \quad (8)$$

pour $\gamma = \beta^2 + (1-\beta)^2$. La preuve de (8) est donnée en annexe. Notons que $-1 < \gamma - 1 < 0$ pour $0 < \beta < 1$. Le ratio des variances avec et sans masquage s'écrit

$$RV_\xi(\bar{y}_{2st} | S,\ S^*) = 1 + m(\gamma - 1)\frac{(\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2})}{\sigma^2_{yu}[1 + (m-1)\rho_{yu}]}. \quad (9)$$

La variance ne variera pas si la permutation est faite de telle manière que $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} = 0$, c'est-à-dire que les corrélations dans les UPE correspondantes soient de sens opposé. Sinon, la variance variera au taux de $m(\gamma - 1) < 0$ pour $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} \neq 0$.

En général, les unités ont tendance à être plus semblables à l'intérieur d'une UPE qu'entre UPE, la valeur de ρ_{yi} étant faible et positive dans de nombreuses populations (par exemple Valliant et coll. 2000, section 8.2.3). Donc, il est plus probable que $\sigma^2_{yi_1}\rho_{yi_1} + \sigma^2_{yi_2}\rho_{yi_2} > 0$, à moins que $\sigma^2_{yi} \doteq 0$ pour tout $i = i_1,\ i_2$ et la variance avec masquage a tendance à être plus faible que la variance sans masquage, c'est-à-dire $RV_\xi(\bar{y}_{2st} | S,\ S^*) < 1$. La figure 2 illustre la variation de l'erreur-type en fonction du CCI sans masquage (ou de référence) ρ_{yu} lorsque la proportion d'unités permutées entre les deux UPE varie. Elle montre que plus le nombre d'unités permutées augmente, plus la variance est modifiée, ce qui indique qu'il faut effectuer la permutation (c'est-à-dire, le masquage des UPE) minimale afin de ne pas induire de biais important dans la variance. En outre, la figure 2 révèle la courbe décroissante en forme de L du ratio des erreurs-types en fonction du CCI, autrement dit indique une surestimation pour les CCI négatifs, mais une sous-estimation pour les CCI positifs. Par conséquent, sous le masquage des UPE, nous pouvons nous attendre à des courbes de type $c_y = 0,7$ (décroissantes, mais aléatoires) ou $c_y = 1,0$ (purement aléatoires) dans la figure 1, le deuxième type représentant les meilleurs résultats qui peuvent être obtenus par masquage minimal. À la section 4, nous proposons une stratégie de masquage des UPE par permutation des USE qui aide à produire une courbe du deuxième type pour les ratios des variances résultants. À la section 5, nous appliquons la stratégie proposée à des données d'enquête artificielles en faisant varier la proportion d'unités permutées.

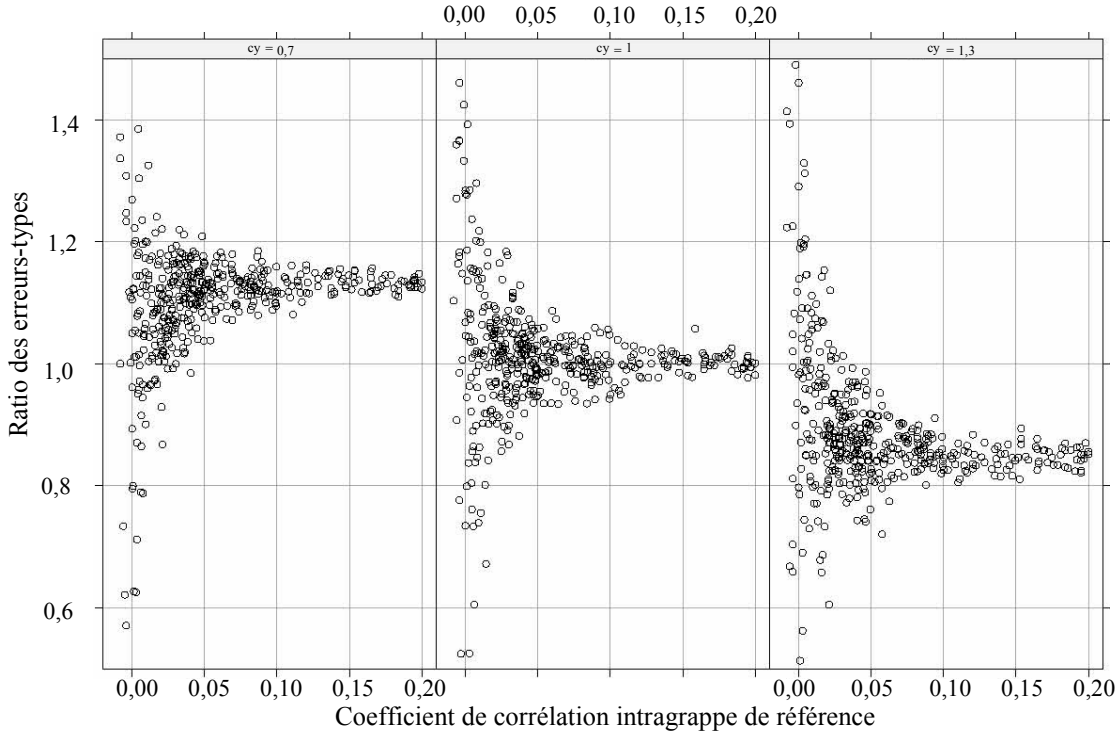


Figure 1 Ratios des erreurs-types avec et sans masquage en fonction du coefficient de corrélation intragrappe original avec variation de l'effet du masquage des UPE sous un modèle $\rho^*_{yU} = c_y\rho_{yU} + e$ avec $e \sim N(0,\ 0,05^2)$ pour $c_y = 0,7,\ 1,0,\ 1,3$ et $m = \bar{M} = S^2_y = 100$

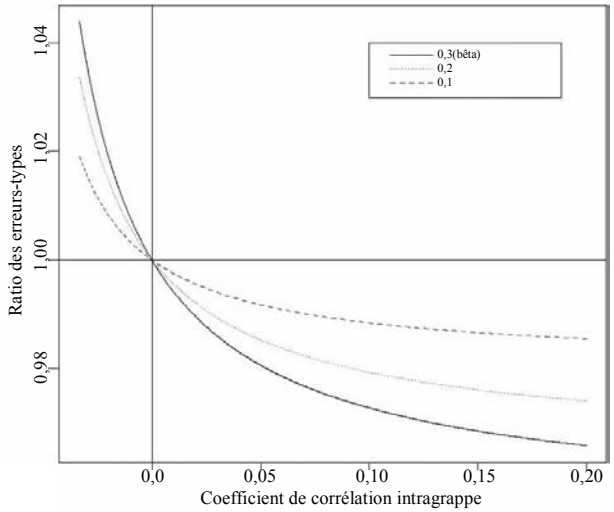


Figure 2 Ratios des erreurs-types avec et sans masquage $\sqrt{RV_{\xi}(\bar{y}_{2st} \mid S, S^*)}$ en fonction du CCI lorsque varie la proportion d'unités permutées provenant de chaque UPE. $\sigma^2_{yu} \equiv 25$, $\rho_{yu} \equiv \rho_{yi}$ pour tout $i, n = 10$, $m = 16$, pour $\beta = (0{,}1,\ 0{,}2,\ 0{,}3)$ et $-0{,}5(m-1)^{-1} \le \rho_{yu} \le 0{,}2$

4. Stratégie de masquage des UPE pour limiter le biais dans l'estimation de la variance

Nombre d'enquêtes à grande échelle comportent plusieurs degrés d'échantillonnage avec probabilités de sélection inégales. Dans de telles circonstances, les unités d'échantillonnage de deuxième degré ou de degré subséquent peuvent être un choix naturel pour la permutation en vue de créer des pseudo-UPE pour des raisons opérationnelles. Par exemple, les diffusions récentes des données de la National Health and Nutrition Examination Survey (NHANES) (Dohrmann et coll. 2005) comprennent les identificateurs de pseudo-UPE construites par permutation des USE entre les UPE originales. À la présente section, nous considérons la permutation des USE en vue de masquer les UPE sous échantillonnage stratifié à plusieurs degrés et son effet sur les estimations de la variance. Nous proposons une stratégie de permutation des USE basée sur la contribution des USE aux estimations des variances.

4.1 USE dans l'estimation de la variance sous échantillonnage stratifié à plusieurs degrés

Supposons qu'une population finie U de M unités est partitionnée en N UPE et que les UPE semblables en ce qui concerne un certain nombre de caractéristiques sont groupées pour former un total de H strates. Supposons aussi que chaque strate est constituée de N_h UPE et que chaque UPE contient N_{hi} USE avec N_{hij} unités finales d'échantillonnage, où $N = \sum_{h=1}^{H} N_h$ et $M = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} N_{hij}$. Supposons que l'échantillonnage de premier degré consiste à sélectionner $n_h = 2$ UPE dans chaque strate, indépendamment entre les strates, et que le deuxième degré et le degré subséquent d'échantillonnage consistent à sélectionner, tour à tour, n_{hi} USE dans chaque UPE échantillonnée (hi) et n_{hij} unités finales dans chaque USE échantillonnée (hij), où $h = 1, ..., H$, $i = 1, ..., n_h$ et $j = 1, ..., n_{hi}$. L'unité finale échantillonnée $(hijk) \in S$ est associée à la valeur observée y_{hijk} de la variable étudiée y et au poids d'échantillonnage w_{hijk}, où $k = 1, ..., n_{hij}$ et S désigne l'ensemble d'indices dans l'échantillon. Le total de population $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} y_{hijk}$ et la taille M sont estimés par $\hat{Y} = \sum_{(hijk) \in S} w_{hijk} y_{hijk}$ et $\hat{M} = \sum_{(hijk) \in S} w_{hijk}$, respectivement. En outre, la moyenne de population $\bar{Y} = Y/M$ est estimée par $\hat{\bar{Y}} = \hat{Y}/\hat{M}$ et son estimateur de variance par développement en série de Taylor (par exemple, Shao et Tu 1995) est donné par

$$v(\hat{\bar{Y}} \mid S) = \sum_{h=1}^{H} \left(\frac{z_{h1} - z_{h2}}{2} \right)^2, \qquad (10)$$

où les $z_{hi} = z_{hi}(y) = \sum_{j=1}^{n_{hi}} \sum_{k=1}^{n_{hij}} 2 w_{hijk} z_{hijk}$ sont les totaux estimés de strate de $z_{hijk} = z_{hijk}(y) = \hat{M}^{-1}(y_{hijk} - \hat{\bar{Y}})$ pour l'UPE (hi).

Dans (10), en écrivant z_{hi} en fonction des unités contenues dans les USE, nous pouvons voir la contribution de ces USE à l'estimation de la variance, donc, faciliter la recherche de stratégies de permutation des USE afin de limiter les biais dans les estimations des variances. Si w_{hij} et $w_{k|hij}$ désignent les poids d'échantillonnage des USE et les poids conditionnels des unités finales d'échantillonnage, respectivement, alors $w_{hijk} = w_{hij} \times w_{k|hij}$. Soit $\hat{N}_{hij} = \sum_{k=1}^{n_{hij}} w_{k|hij}$ et $\hat{\bar{Y}}_{hij} = \hat{N}_{hij}^{-1} \sum_{k=1}^{n_{hij}} w_{k|hij} y_{hijk}$ les estimations de la taille et de la moyenne d'échantillon de l'USE (hij), respectivement. Dans (10), les quantités z_{hi} peuvent maintenant s'écrire

$$z_{hi} = \sum_{j=1}^{n_{hi}} 2 w_{hij} z_{hij}, \qquad (11)$$

où

$$z_{hij} = \sum_{k=1}^{n_{hij}} w_{k|hij} z_{hijk} = \hat{M}^{-1} \hat{N}_{hij} (\hat{\bar{Y}}_{hij} - \hat{\bar{Y}}).$$

Il est évident, si l'on examine (10) et (11), que la contribution des USE échantillonnées à l'estimation de la variance a lieu par la voie des trois composantes $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ de l'USE (hij). À la section 4.2, nous examinerons de près l'effet du masquage des UPE par permutation des USE sur les estimations des variances.

4.2 Effet de la permutation des USE sur les estimations de la variance

Nous supposons maintenant que deux USE $(h_a\, i_a\, j_a)$ et $(h_b\, i_b\, j_b)$ sont permutées entre deux UPE $(h_a\, i_a) \neq (h_b\, i_b)$.

Alors, l'estimation de la variance avec masquage peut, en partant de (10), s'écrire

$$v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S) + \sum_{h \in \{h_a, h_b\}} \left[\left(\frac{z^*_{h1} - z^*_{h2}}{2} \right)^2 - \left(\frac{z_{h1} - z_{h2}}{2} \right)^2 \right], \quad (12)$$

où z^*_{hi} désigne la quantité z_{hi} dans (11) avec l'ensemble d'indices dans l'échantillon S^* altéré à cause de la permutation. Désignons par i'_a et i_b les autres UPE dans les strates h_a et h_b, respectivement, et définissons $z_{hi(j)} = z_{hi} - z_{hi} - 2w_{hij} z_{hij} = \sum_{l \neq j} 2w_{hil} z_{hil}$. Alors, (12) peut s'écrire

$$v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S) + e_0(y) g_0(y), \quad (13)$$

où $e_0(y) = 2(w_{h_a i_a j_a} z_{h_a i_a j_a} - w_{h_b i_b j_b} z_{h_b i_b j_b})$ est la différence de quantité $2w_{hij} z_{hij} = 2w_{hij} \hat{M}^{-1} (\hat{\bar{Y}}_{hij} - \hat{\bar{Y}})$ entre les deux USE qui seront permutées et

$$g_0(y) = \begin{cases} [z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)}] & \text{si } h_a = h_b, \\ 2^{-1}[(z_{h_a i'_a} - z_{h_a i_a (j_a)}) - (z_{h_b i'_b} - z_{h_b i_b (j_b)})] & \text{si } h_a \neq h_b, \end{cases}$$

est une fonction de $2w_{hij} z_{hij}$ des USE qui seront gardées dans les UPE originales. Notons que, pour $h_a = h_b$, r_0 peut aussi s'exprimer $g_0 = 2^{-1}[(z_{h_b i_b (j_b)} - z_{h_a i_a (j_a)}) - (z_{h_a i_a (j_a)} - z_{h_b i_b (j_b)})]$ ce qui montre que l'effet de la permutation des USE sur l'estimation de la variance sera négligeable si les deux USE permutées sont appariées de manière que le produit de $e_0(y)$ et $g_0(y)$ soit proche de zéro. Autrement dit, la variation de l'estimation de la variance sous permutation des USE peut être contrôlée si l'on forme une paire de segments tenant compte des trois composantes $\{w_{hij}, \hat{N}_{hij}, \hat{\bar{Y}}_{hij}\}$ de manière à minimiser $e_0(y) \times g_0(y)$ similairement au cas sous échantillonnage en grappes PPT à un degré de la section 2.2.

De plus, en écrivant

$$g_0 = g(e_0) = \begin{cases} (z_{h_a 1} - z_{h_a 2}) - e_0 & \text{si } h_a = h_b, \\ 2^{-1}\left[(z_{h_a 1} - z_{h_a 2}) - (z_{h_b 1} - z_{h_b 2}) - e_0\right] & \text{si } h_a \neq h_b, \end{cases}$$

nous pouvons exprimer (13) comme une fonction quadratique de $e_0(y)$ sachant $\{z_{hi} : h = \{h_a, h_b\}, i = 1, 2\}$. Pour $h_a = h_b$, nous pouvons montrer que $v(\hat{\bar{Y}} \mid S^*) > v(\hat{\bar{Y}} \mid S)$ uniquement pour e_0 compris entre zéro et $(z_{h_a 1} - z_{h_b 2})$. Quand $z_{h_a 1} - z_{h_b 2} \doteq 0$, il pourrait être plus probable que $v(\hat{\bar{Y}} \mid S^*) = v(\hat{\bar{Y}} \mid S)$. Un argument comparable peut être formulé pour $h_a \neq h_b$.

4.3 Permutation séquentielle d'USE sous caractéristiques d'appariements multiples

Supposons que l'échantillon contient un total de n_J USE et que seulement R d'entre elles sont choisies pour former des paires pour la permutation, où $n_J = \sum_{h=1}^{H} \sum_{i=1}^{n_h} n_{hi}$ et $1 \leq R < n_J$. Supposons qu'un nombre fixe de R USE est choisi conformément à une certaine notion de compromis risque-utilité. Voir, par exemple, Gomatam, Karr et Sanil (2005) pour une discussion apparentée concernant la permutation de données. En outre, supposons que leur ordre séquentiel pour le processus d'appariement est donné par $j_1, j_2, ..., j_R$. Par exemple, pour commencer, toutes les paires possibles sont formées pour chacune des R USE et la meilleure d'entre elles est choisie en se fondant sur une mesure de distance, telle que (12). L'ordre des R USE pour le (principal) processus d'appariement est alors déterminé en fonction de l'ordre croissant des distances des R meilleures paires.

Soit S^{r-1} l'ensemble modifié d'indices dans l'échantillon après que la $(r-1)^{\text{e}}$ paire d'USE a été formée et permutée, où $r = 1, ..., R$ et $S^0 \equiv S$. Soit $S^{r-1}_{(j)}$ l'ensemble d'indices dans l'échantillon quand les USE j_r et j sont permutées pour S^{r-1}. Alors, la variation de l'estimation de la variance causée par la permutation de la r^{e} USE j_r et toute autre USE non comprise dans les $(r-1)$ appariements précédents peut s'écrire sous la forme

$$\begin{aligned} \delta_r(y, j) &= v(\hat{\bar{Y}} \mid S^{r-1}_{(j)}) - v(\hat{\bar{Y}} \mid S^{r-1}) \\ &= e_{r-1}(y, r)\, g_{r-1}(y, r), \end{aligned} \quad (14)$$

où $e_{r-1}(y, r)$ et $g_{r-1}(y, r)$ sont définis de la même façon que dans (13), mais avec S^{r-1} et $S^{r-1}_{(j)}$. Manifestement, le choix de la meilleure paire pour la r^{e} USE dépend des $(r-1)$ appariements précédents et, donc, le processus d'appariement doit être considéré comme étant séquentiel. Notons que les USE qui ont été appariées et permutées au moment des appariements précédents doivent être exclues de la recherche courante.

En outre, plusieurs caractéristiques peut être prises en considération pour l'appariement, dans l'espoir qu'elles seront reliées à de nombreuses autres variables étudiées, de façon à minimiser le biais dans l'estimation connexe de la variance. Supposons que q caractéristiques d'appariement sont choisies avec soin, disons $\mathbf{x} = (x_1, x_2, ..., x_q)'$ (par exemple, Dohrmann et coll. 2005, pour une discussion à ce sujet). Pour déterminer la distance entre les USE j_r et j, nous pouvons considérer toute mesure de distance de la forme

$$D_r(j \mid \mathbf{x}) = \sum_{l=1}^{q} c_l \left| v[\hat{\bar{X}}_l \mid S^{r-1}_{(j)}] - v[\hat{\bar{X}}_l \mid S^{r-1}] \right| \quad (15)$$

ou

$$\Delta_r(j \mid \mathbf{x}) = \sum_{l=1}^{q} c_l \left| v[\hat{\bar{X}}_l \mid S^{r-1}_{(j)}] - v[\hat{\bar{X}}_l \mid S] \right| \quad (16)$$

moyennant tout choix raisonnable de coefficients positifs c_l. Par exemple, $c_l \equiv 1$ tient compte simplement de l'écart

absolu entre les estimations de la variance de $\hat{X}_l$, $c_l = v(\hat{\bar{X}}_l \mid S)^{-1}$, de l'écart absolu entre les estimations de la variance comparativement aux estimations originales de la variance et $c_l = X_l^{-1}$, de l'écart absolu entre les estimations de la variance relative. La première mesure de distance (15) rend compte de la variation de l'estimation de la variance due à la permutation de segments de la r^{e} paire. La deuxième mesure de distance (16) tient compte de l'effet cumulatif des permutations pour les r paires de segments.

Nous pouvons imposer des contraintes d'appariement, par exemple, pour empêcher l'appariement d'USE provenant de la même UPE et pour appliquer un seuil de la proportion d'USE provenant de chaque UPE qui seront permutées (Lu 2004). Soit $J_A = \{j_1, \ldots, j_R\}$ l'ensemble d'indices des R USE prises en considération pour former les paires à permuter et soit J_B toutes les USE possibles qui peuvent être appariées en satisfaisant un ensemble donné de contraintes d'appariement. Pour simplifier, considérons que l'appariement des USE n'est pas permis à l'intérieur de J_A, c'est-à-dire $J_A \cap J_B = \varnothing$. Si $D_r^*(j \mid \mathbf{x})$ désigne la mesure de distance choisie pour les USE j_r et j, un algorithme de permutation séquentielle des USE en vue de limiter les biais dans les estimations de la variance peut prendre la forme suivante :

Étape 1. Poser que $r = 1$, $J_A^r = J_A$ et $J_B^r = J_B$;

Étape 2. Pour chacune des $(R - r + 1)$ USE dans J_A^r, calculer $D_r^*(\cdot \mid \mathbf{x})$ pour toutes les USE dans J_B^r ;

Étape 3. Choisir le meilleur appariement ayant la distance $D_r^*(\cdot \mid \mathbf{x})$ la plus faible, autrement dit, trouver j_r' telle que $D_r^*(j_r' \mid \mathbf{x}) = \min_{j \in J_B^r} D_r^*(j \mid \mathbf{x})$;

Étape 4. Poser que $r = r + 1$ et retirer la paire choisie du groupe d'unités sur lequel porte la recherche, autrement dit, poser que $J_A^r = J_A^{r-1} \setminus \{j_r\}$ et mettre à jour J_B^r en conséquence, où $J_B^r \subseteq J_B^{r-1} \setminus \{j_r'\}$;

Étape 5. Si $r = R + 1$, arrêter ; sinon, répéter les étapes 2 à 4.

Cette méthode d'appariement (ou de permutation) des USE consiste fondamentalement à rechercher la paire qui, à l'étape du r^{e} appariement, est la meilleure au sens de la minimisation de la variation des estimations de la variance due à la permutation des USE correspondantes. Si le nombre d'USE est grand, cette méthode produit un diagramme de dispersion semblable à celui illustré pour $c_y = 1{,}0$ à la figure 1 (c'est-à-dire, une perturbation aléatoire avec une courbe en forme d'entonnoir).

Le choix de critères d'optimalité plus perfectionnés en ce qui concerne $\{c_l : l = 1, \ldots, q\}$ pourrait aider à améliorer la méthode susmentionnée en vue de réduire l'ampleur de ce genre de perturbation aléatoire dans les estimations de la variance. En outre, en utilisant des techniques multivariées, telles que l'analyse en composantes principales, pour établir certains types de scores (par exemple, un ou plusieurs axes factoriels (composantes principales)) à partir d'un grand nombre de caractéristiques continues, il serait peut-être possible de réduire davantage l'importance de cette perturbation aléatoire dans les estimations de la variance. À la section 5, nous donnons des exemples en ce qui concerne la permutation des USE.

5. Exemples

5.1 Travaux antérieurs

Dans le cas d'un plan d'échantillonnage sans stratification, mais comportant un petit nombre d'UPE, Dohrmann et coll. (2002) ont examiné diverses méthodes de partitionnement des UPE en pseudo-UPE afin d'appliquer la méthode du jackknife avec suppression d'une unité pour estimer la variance. Leur idée fondamentale était de doubler le nombre d'UPE masquées en gardant les UPE fractionnées comme des UPE masquées distinctes, espérant ainsi réduire le risque de divulgation des données grâce à la rupture du lien entre les UPE réelles et masquées. Leur étude empirique a révélé, dans les estimations résultantes de la variance pour les variables présentant des effets de plan importants, des courbes de sous-estimation appréciable qui ressemblent au graphique de $c_y = 0{,}7$ à la figure 1. Soit S et $S^{\dagger}$ les ensembles d'indices dans l'échantillon sans masquage et avec masquage, respectivement. Soit w_{ij} le poids d'échantillonnage et soit y_{ij} la valeur observée de y pour la j^{e} unité échantillonnée dans l'UPE i. Pour expliquer les courbes de sous-estimation observées, Dohrmann et coll. (2005) ont établi la relation suivante

$$v(\hat{\bar{Y}} \mid S^{\dagger}) = \frac{n-1}{2n-1} v(\hat{\bar{Y}} \mid S) + \frac{1}{n(2n-1)} \sum_{i=1}^{n} (z_{1,i} - z_{2,i})^2,$$

où les $z_{g,i} = \sum_{j \in S_{g,i}} 2 w_{ij} z_{ij}$ sont les totaux dans les fractions d'UPE de $z_{ij} = (\sum_{ij} w_{ij})^{-1} (y_{ij} - \sum_{ij} w_{ij} y_{ij} / \sum_{ij} w_{ij})$ et $S_{u,i}$ sont les ensembles d'indices de la u^{e} fraction de l'UPE i pour $i = 1, \ldots, n$ et $u = 1, 2$. Cette expression indique que l'estimation de la variance résultante est égale à environ la moitié de celle obtenue sans masquage à laquelle s'ajoute une valeur positive reflétant les totaux entre les fractions d'UPE des z_{ij} dans les UPE. Si les ensembles $S_{u,i}$ sont formés de manière que $z_{1,i} \doteq z_{2,i}$, cette méthode de partitionnement des UPE donne lieu à environ la moitié de l'estimation de la variance sans masquage et l'estimation de la variance avec masquage doit donc être doublée pour s'approcher de la valeur sans masquage.

Dans le cas du plan à deux UPE par strate, Dohrmann et coll. (2005) ont envisagé une autre approche dans laquelle les pseudo-UPE sont construites par permutation des USE

entre les UPE. Comme nous le mentionnons à la section 1, cette approche peut être considérée comme la division des UPE en une ou plusieurs fractions et la recombinaison de ces fractions pour construire des pseudo-UPE dans lesquelles les fractions d'UPE sont permutées. Pour simplifier, nous supposons que chaque UPE est divisée en deux fractions $S_{1,hi}$ et $S_{2,hi}$. Si la partition est telle que $S_{h1}^{\dagger} = S_{1,h1} \cup S_{1,h2}$ et $S_{h2}^{\dagger} = S_{2,h1} \cup S_{2,h2}$, l'estimation de la variance avec masquage peut s'écrire

$$v(\hat{\bar{Y}} \mid S^{\dagger}) = v(\hat{\bar{Y}} \mid S) + \sum_{h=1}^{H} e_h(y)\, g_h(y), \qquad (17)$$

où $e_h(y) = z_{1,h2} - z_{2,h1}$ est la différence entre les totaux dans les fractions d'UPE $S_{1,h2}$ et $S_{2,h1}$ à permuter et $g_h(y) = z_{1,h1} - z_{2,h2}$ est la différence entre les totaux dans les fractions d'UPE $S_{1,h1}$ et $S_{2,h2}$ à garder dans les UPE originales. La preuve de (17) est donnée en annexe. L'équation (17) indique qu'une stratégie semblable pour le partitionnement des UPE aiderait à préserver la grandeur de l'estimation originale de la variance. Dohrmann et coll. (2005) ont adopté une méthode probabiliste de couplage d'enregistrements (Fellegi et Sunter 1969) pour former des paires d'USE pour la permutation dont les moyennes $\hat{\bar{Y}}_{hij}$ sont semblables pour plusieurs caractéristiques, dans l'espoir que, dans (17), les termes e_h soient tous proches de zéro. Dohrmann et coll. (2005) ont démontré que la méthode de recombinaison des UPE permet dans une certaine mesure de réduire les biais dans les estimations de la variance et les courbes de sous-estimation résultantes comparativement à la méthode de partitionnement des UPE appliquée dans Dohrmann et coll. (2002). Afin de rendre la méthode plus rapide et plus souple, Lu (2004) a élaboré des algorithmes de permutation des USE basés sur l'évaluation séquentielle de mesures de distance entre les moyennes d'USE, sans tenir compte directement de l'incidence des permutations successives sur le biais de l'estimation de la variance. Comme nous l'avons exposé à la section 4.2, l'effet de la permutation des USE sur les estimations de la variance peut être réduit davantage en tenant compte directement de la contribution des USE à ces estimations. À la section suivante, nous appliquons deux stratégies, celle décrite par Dohrmann et coll. (2005) et celle proposée à la section 4.2, à des données artificielles provenant d'une enquête complexe.

5.2 Exemple de données

Afin d'illustrer l'effet du masquage des UPE sur les estimations de la variance des moyennes d'échantillon, nous avons utilisé des données d'enquête réelles en apportant certaines modifications artificielles au fichier à grande diffusion (FGD) de données sur les objectifs en matière de santé pour l'an 2000 de la National Health Interview Survey (NHIS) de 1993. La NHIS est une enquête-ménages annuelle avec interview sur la santé menée auprès de la population civile non placée en établissement des États-Unis. La NHIS est réalisée selon un plan d'échantillonnage à plusieurs degrés stratifié typique dans lequel les UPE de premier degré correspondent aux comtés ou aux régions métropolitaines et les USE de deuxième degré correspondent à des segments (c'est-à-dire, un petit nombre de ménages dans une petite région géographique) à l'intérieur des UPE échantillonnées. Le questionnaire particulier sur le thème de l'an 2000 a été administré à un seul adulte échantillonné par famille durant la deuxième moitié de 1993. Les données de la NHIS utilisées ici et la documentation les concernant peuvent être consultées dans National Center for Health Statistics (1994) ou sur le site Web du National Center for Health Statistics des Centers for Disease Control des États-Unis (http://www.cdc.gov/nchs/about/major/nhis/quest_data_related_doc.htm).

Ce FGD contient les identificateurs de strate et d'UPE, ainsi que les poids finaux des personnes échantillonnées pour l'estimation de la variance. Dans notre exemple, nous n'avons utilisé que dix strates, mais avons limité le nombre d'UPE à deux par strate. Deux des strates sélectionnées, 110 et 520, ont été limitées à leurs deux plus grandes UPE, (181, 410) et (048, 233) respectivement, et les huit autres strates, 102, 142, 192, 211, 261, 300, 561 et 571, ne contiennent que deux UPE. Le FGD contient aussi les identificateurs des USE, mais non leurs poids d'échantillonnage. Pour produire les poids d'échantillonnage des USE, w_{hij}, nous nous sommes servis d'un modèle à deux facteurs avec effets aléatoires emboîtés pour ajuster $\log w_{hijk} = \log w_{hij} + \log w_{k|hij}$ de manière que $\log w_{hij} = \mu + \alpha_{hi} + \beta_{j(hi)}$ et $\log w_{k|hij} = \varepsilon_{k(hij)}$, où μ est une valeur commune, α_{hi} est l'effet aléatoire de l'UPE (hi), $\beta_{j(hi)}$ est l'effet aléatoire de l'USE j emboîtée dans l'UPE (hi) et $\varepsilon_{k(hij)}$ est l'effet aléatoire de la personne échantillonnée k dans l'USE (hij). Nous avons restreint notre étude seulement aux USE comptant cinq personnes échantillonnées ou plus, ce qui nous a donné un total de 293 USE pour l'analyse. La décomposition résultante des poids $(w_{hij}, w_{k|hij})$ peut comprendre d'éventuelles spécifications incorrectes du modèle, mais elle suffit pour les besoins de l'illustration, puisque les w_{hij} ainsi que les $w_{k|hij}$ sont tous positifs sous le modèle. Pour obtenir des paires d'USE pour la permutation, nous avons utilisé six variables sociodémographiques, désignées par $x_1, x_2, \ldots, x_6$. Elles sont énumérées au tableau 1 avec leur description, leur définition, leur moyenne globale d'échantillon et la racine carrée de leur effet de plan (« design factor » ou deft en abrégé). Les moyennes d'échantillon de ces variables varient de 0,05 à 0,63 et les valeurs deft, de 1,285 à 8,511.

Nous avons appliqué les deux stratégies d'appariement des USE discutées aux sections 4 et 5.1, respectivement. La

première consiste à employer une mesure de distance (15) pour toute paire d'USE (r_a, r_b) avec $c_l \equiv 1$ pour tout $l = 1, ..., 6$. Soit S^{r-1} et S^r les deux ensembles d'indices dans l'échantillon après les $(r-1)^{\text{e}}$ et r^{e} permutations, respectivement. Alors, la distance de la r^{e} paire appariée de la première stratégie (variance-appariement) s'écrit

$$D_r(v|\mathbf{x}) = \sum_{l=1}^{6} \left| v(\hat{\bar{X}}_l|S^r) - v(\hat{\bar{X}}_l|S^{r-1}) \right|,$$

où $v(\hat{\bar{X}}_l|S^r)$ et $v(\hat{\bar{X}}_l|S^{r-1})$ représentent les estimations de la variance de $\hat{\bar{X}}_l$ pour la l^{e} caractéristique d'appariement avec S^r et S^{r-1}, respectivement. Le biais dans les estimations de la variance qui résulte de la permutation de la r^{e} paire appariée est d'autant plus faible que la distance est courte. La deuxième stratégie, proposée par Dohrmann et coll. (2005), consiste à apparier les USE dont les moyennes d'échantillon sont semblables pour les six caractéristiques d'appariement. Cette stratégie (moyenne-appariement) définit la distance de la r^{e} paire appariée comme étant :

$$d_r(\mu \mid \mathbf{x}) = \sum_{l=1}^{6} \left| \hat{\bar{X}}_{l,r_a} - \hat{\bar{X}}_{l,r_b} \right|,$$

où $\hat{\bar{X}}_{l,r_i}$ représente la moyenne d'échantillon de l'USE r_i $(i = a, b)$ pour la caractéristique d'appariement x_l.

Tableau 1
Variables utilisées pour l'appariement

Variable	Description	Définition	Moyenne d'échantillon	Valeur deft
x_1	Sexe masculin	SEX=1	0,49	1,285
x_2	Hispanicité	HISPANIC	0,14	8,511
		= 00, 01, ..., 08	0,14	8,511
x_3	Couple marié	MARSTAT = 1,2	0,63	3,209
x_4	Études collégiales ou de niveau plus élevé	EDUCR = 4, 5, 6	0,45	2,902
x_5	Revenu familial élevé (50 000 $ ou plus)	INCFAMR = 8	0,23	3,558
x_6	L'air ambiant dans le ménage a-t-il été testé pour la présence de radon ?	TESTRDN = 1	0,05	2,191

Le tableau 2 donne les ratios des erreurs-types pour les six caractéristiques d'appariement à chaque étape d'appariement dans l'ordre séquentiel pour chaque stratégie avec 18 paires permutées (environ 12 % de la totalité des USE dans les données). La première stratégie, présentée dans la partie de gauche du tableau, donne un intervalle modéré, mais légèrement croissant, de variation des ratios des erreurs-types sur la série des 18 paires sélectionnées pour la permutation. La deuxième stratégie, présentée dans la partie de droite du tableau, produit un intervalle assez vaste de variation des ratios des erreurs-types sur la série, les changements étant très grands pour la treizième paire et les suivantes de la série de permutations. Bien que ces stratégies tendent toutes deux à perdre le contrôle sur le biais dans les estimations de la variance pour les ordres élevés de la série de permutations, la première permet de relativement bien maîtriser le biais dans les estimations de la variance pour un nombre assez grand de paires à permuter.

À la figure 3, les ratios des erreurs-types sont tracés en fonction des valeurs deft pour les deux stratégies en faisant varier le nombre d'USE permutées. Les trois ensembles comprennent 6 (4 %), 12 (8 %) et 18 (12 %) paires d'USE (pourcentage d'USE intervenant dans les permutations), respectivement. Chaque graphique comprend deux ensembles de caractéristiques, six caractéristiques d'appariement marquées par les chiffres correspondant à ceux énumérés au tableau 1 et 92 caractéristiques marquées d'un × qui ne sont pas utilisées dans l'appariement. Dans le scénario où 4 % seulement des USE sont permutées, la différence entre les deux stratégies est négligeable pour les deux ensembles de caractéristiques. Cependant, à mesure que le pourcentage d'USE permutées augmente, la perturbation dans les estimations de la variance s'accroît pour les deux stratégies et les deux ensembles de caractéristiques. Ce résultat indique que le pourcentage de permutations effectuées devrait être faible, ce qui renforce les résultats de la section 3. En outre, les ratios des erreurs-types sont regroupés plus près de la ligne de valeur un (c'est-à-dire petit biais dans les estimations de la variance avec masquage) dans la première stratégie que dans la seconde. Cette dernière produit une courbe assez fortement décroissante en fonction des valeurs deft, même pour les six caractéristiques d'appariement. Autrement dit, la stratégie moyenne-appariement semble avoir un effet plus prononcé pour les variables utilisées pour l'appariement.

Tableau 2
Ratios des erreurs-types selon la série de permutations : comparaison de deux critères d'appariement avec permutation de 12 % (18 paires) d'USE

Série de permutations	Variance-appariement						Moyenne-appariement					
	x_1	x_2	x_3	x_4	x_5	x_6	x_1	x_2	x_3	x_4	x_5	x_6
1	0,999	1,000	1,000	0,998	1,000	0,998	0,998	1,000	1,000	1,002	1,002	1,001
2	1,002	1,000	1,000	1,000	1,000	0,996	0,999	1,000	1,000	1,002	1,001	1,001
3	1,004	1,000	1,000	1,001	1,000	0,996	0,998	1,000	0,999	0,997	0,994	1,001
4	1,012	1,001	0,999	1,000	1,000	0,989	1,025	1,000	0,999	1,001	0,994	1,016
5	1,009	1,001	1,000	0,998	1,001	0,988	1,021	1,004	0,964	0,968	0,951	1,013
6	1,007	1,000	1,000	1,000	1,003	0,988	1,020	1,004	0,964	0,968	0,955	1,015
7	1,011	1,000	1,000	1,002	1,003	1,008	1,020	1,004	0,964	0,970	0,954	1,017
8	1,016	1,000	0,997	1,002	1,003	1,026	1,022	0,998	0,957	0,963	0,964	1,005
9	1,009	0,998	1,000	1,002	1,003	1,020	1,021	0,997	0,955	0,965	0,982	1,034
10	1,007	0,996	1,001	1,010	1,006	1,014	1,019	0,997	0,931	0,960	0,972	1,033
11	1,014	0,994	1,005	1,010	1,001	1,012	1,020	0,995	0,946	0,953	0,989	1,034
12	1,029	0,995	1,003	1,013	1,011	1,036	1,021	0,991	0,946	0,953	0,987	1,035
13	1,064	0,992	1,003	1,001	1,008	1,047	1,035	0,990	0,946	0,932	0,967	1,114
14	1,008	0,991	1,000	1,007	1,022	1,044	1,031	0,955	0,946	0,929	0,952	1,103
15	1,042	0,988	0,984	1,017	1,015	1,044	1,052	0,955	0,946	0,922	0,952	1,124
16	1,012	0,982	0,986	1,024	1,041	1,042	1,107	0,939	0,936	0,920	0,942	1,128
17	0,987	0,978	1,000	1,016	1,009	1,021	1,107	0,878	0,936	0,927	0,935	1,123
18	1,029	0,943	1,000	0,970	0,947	1,042	1,014	0,538	0,946	0,841	0,945	1,106

Voir le tableau 1 pour la description des six caractéristiques d'appariement (x_1, ..., x_6).

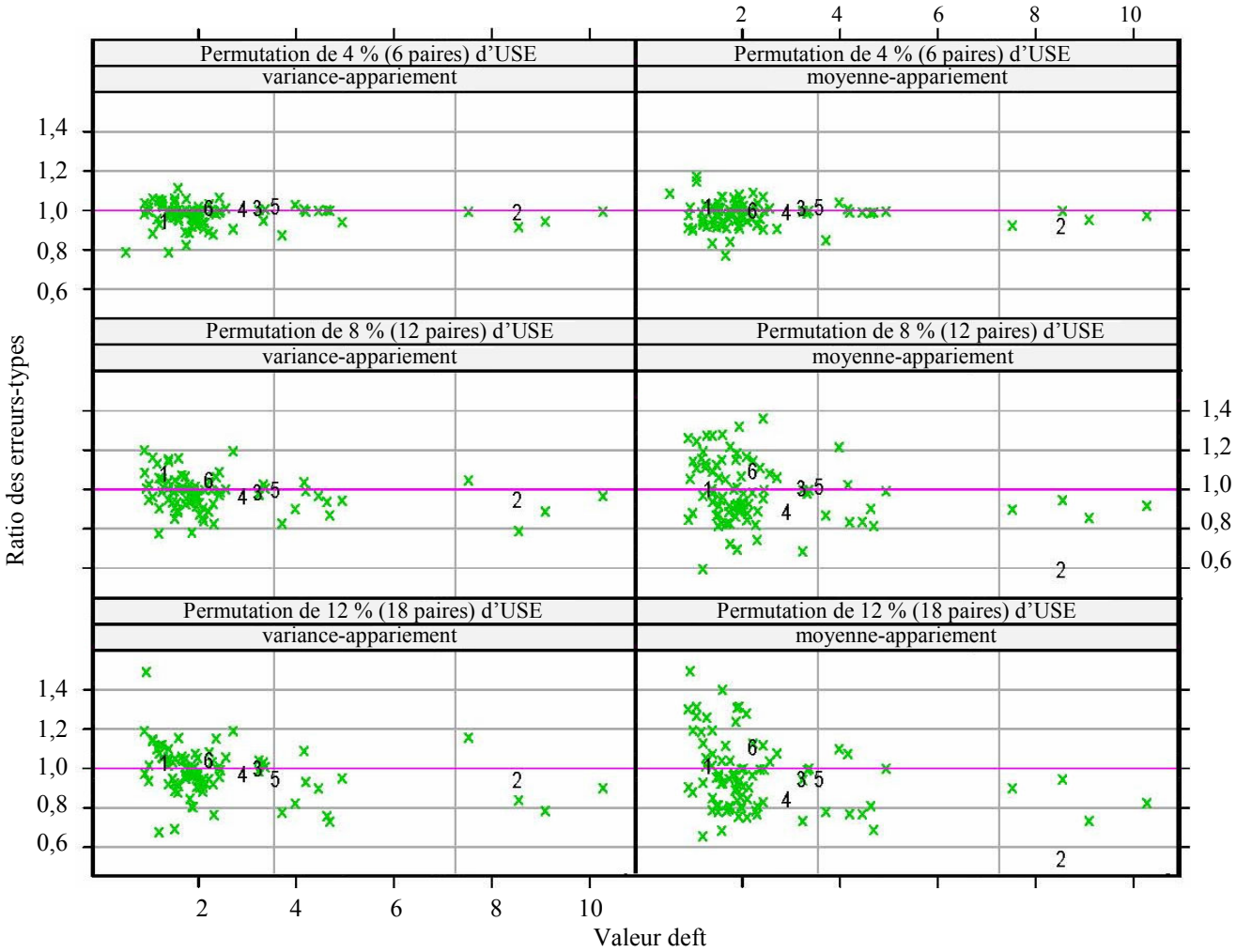


Figure 3 Rapport des erreurs-types en fonction des valeurs deft de référence. Six chiffres représentent les points pour les caractéristiques d'appariement correspondantes et les × représentent ceux des 92 caractéristiques qui n'ont pas été utilisées pour l'appariement

6. Discussion

Dans le présent article, nous avons étudié l'effet du masquage des UPE sur les estimations de la variance dans les enquêtes complexes. Manifestement, ce masquage fausse la structure de mise en grappes du plan d'échantillonnage original, ce qui peut induire des biais systématiques dans l'analyse des données résultantes, comme l'illustrent les sections 2, 3 et 5.2. La stratégie de masquage des UPE proposée à la section 4 peut aider à réduire ces biais, mais laisse néanmoins une perturbation aléatoire dans l'estimation de la variance et, donc, une perte d'efficacité inférentielle. Il serait intéressant d'étudier l'effet du masquage des UPE sur d'autres types d'analyse de données complexes, comme les analyses par régression et les analyses multivariées. Bien que le masquage des UPE puisse offrir un moyen de contrôler la divulgation, le degré de masquage doit être minimal afin de limiter les biais résultants dans les estimations de la variance, comme il est discuté aux sections 3 et 5.2.

En outre, la réduction du risque d'identification attribuable au masquage des USE pourrait être mieux comprise en exprimant la distance entre la moyenne des UPE échantillonnées masquées et la moyenne des UPE dans la population de la façon suivante :

$$\hat{\bar{Y}}_{hi|S^\dagger} - \bar{Y}_{hi|U} = (\hat{\bar{Y}}_{hi|S^\dagger} - \hat{\bar{Y}}_{hi|S}) + (\hat{\bar{Y}}_{hi|S} - \bar{Y}_{hi|U}), \qquad (18)$$

où $\hat{\bar{Y}}_{hi|S^\dagger}$ et $\hat{\bar{Y}}_{hi|S}$ désignent les moyennes des UPE masquées et non masquées dans l'échantillon, respectivement, et $\bar{Y}_{hi|U}$ désigne la moyenne des UPE dans la population auxquelles pourrait avoir accès un intrus (c'est-à-dire un utilisateur de données malveillant) à partir de sources externes, telles que les données de recensement. Il est facile de montrer qu'en général, le premier terme du deuxième membre de (18) n'est pas nul en cas de masquage des UPE. Les termes $\hat{\bar{Y}}_{hi|S^\dagger}$ et $\bar{Y}_{hi|U}$, conjugués à une variation dans l'échantillon non négligeable du deuxième terme du membre de (18), ne sont jamais égaux, sauf dans des cas très rares. Dohrmann et coll. (2005) comparent $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ de l'échantillon et $\{\bar{Y}_{hi|U}\}$ de la population au moyen d'un diagramme arborescent élégant pour démontrer à quel point un intrus aurait de la difficulté à identifier une UPE échantillonnée dans les fichiers de données à grande diffusion compte tenu de deux aspects, à savoir que 1) peu de paires $(\hat{\bar{Y}}_{hi|S^\dagger}, \bar{Y}_{hi|U})$ sont proches l'une de l'autre et 2) de nombreuses UPE non échantillonnées ont des valeurs de population semblables à $\{\hat{\bar{Y}}_{hi|S^\dagger}\}$ ou à $\{\bar{Y}_{hi|U}\}$ des UPE échantillonnées. Certaines formes de mesure probabiliste pourraient être intéressantes pour évaluer la réduction du risque d'identification (par exemple, Eltinge 1999), mais elles dépassent le cadre du présent article. La stratégie de masquage proposée a été appliquée aux données diffusées de la National Health and Nutrition Examination Survey (NHANES) de 2003-2004 (Park, Dohrmann, Montaquila, Mohadjer et Curtin 2006).

Annexe

Preuves

Preuve de l'équation (1)

D'après Park et Lee (2004, section 4.2),

$$\begin{aligned} V(\hat{\bar{Y}} \mid S) &= \frac{1}{m} S_y^2 \times \mathrm{Deft}^2(\hat{\bar{Y}} \mid S) \\ &\doteq \frac{1}{m} S_y^2 [1 + (\bar{M} - 1)\rho_{yU}] \\ &\quad + \frac{1}{mN} \sum_{i=1}^N \frac{M_i}{p_i}\left(\frac{M_i}{p_i} - p_i\right)(\bar{Y}_i - \bar{Y})^2 \\ &\doteq \frac{1}{m(N-1)} \sum_{i=1}^N M_i (\bar{Y}_i - \bar{Y})^2 \\ &\quad + \frac{1}{mN} \sum_{i=1}^N \frac{M_i}{p_i}\left(\frac{M_i}{M} - p_i\right)(\bar{Y}_i - \bar{Y})^2, \\ &\doteq \frac{1}{mMN} \sum_{i=1}^N \frac{M_i^2}{p_i} (\bar{Y} - \bar{Y})^2, \end{aligned}$$

où $\mathrm{Deft}^2(\hat{\bar{Y}} \mid S)$ représente l'effet de plan de $\hat{\bar{Y}}$ pour un S donné, les deuxième et dernière approximations découlent de $(N-1)/N \doteq 1$ et la troisième équation, de $S_y^2[1 + (\bar{M} - 1)\rho_{yU}] \doteq (N-1)^{-1}\sum_{i=1}^N M_i(\bar{Y}_i - \bar{Y})^2$, ce qui complète la preuve.

Preuve de l'équation (8)

Par définition, la variance du total des UPE échantillonnées $y_i = \sum_j y_{ij}$ est $V_\xi(y_i \mid S) = m\sigma_{yi}^2 + m(m-1)\sigma_{yi}^2\rho_{yi}$ pour $i = 1, ..., n$. Supposons que les $\{y_{ij} : j = m(\beta - 1) + 1, ..., n\}$ pour les deux UPE $i = a$ et b sont permutées. Alors, la variance de ces deux UPE devient $V_\xi(y_a \mid S^*) = m(1-\beta)\sigma_a^2 + m(1-\beta)[m(1-\beta)-1]\sigma_a^2\rho_a + m\beta\sigma_b^2 + m\beta(m\beta - 1)\sigma_b^2\rho_b$ et $V_\xi(y_b \mid S^*)$ donnée par la même expression en permutant les indices a et b. Puisque $m(1-\beta)[m(1-\beta)-1] + m\beta(m\beta - 1) = m(m\gamma - 1)$, la preuve est complétée en observant que

$$\begin{aligned} (nm)^2 V_\xi(\bar{y}_{2st} \mid S^*) &= \sum_{i \neq a,b}^n [m\sigma_{yi}^2 + m(m-1)\sigma_{yi}^2\rho_{yi}] \\ &\quad + \sum_{i=a,b} [m\sigma_{yi}^2 + m(m\gamma - 1)\sigma_{yi}^2\rho_{yi}] \\ &= (nm)^2 V_\xi(\bar{y}_{2st} \mid S) \\ &\quad + \sum_{i=a,b} m^2(\gamma - 1)\sigma_{yi}^2\rho_{yi}. \end{aligned}$$

Preuve de l'équation (13)

Supposons que deux UPE (h1) et (k1) provenant de deux strates distinctes $h \neq k$ sont reconstruites en

permutant chacune de leur USE, $(h1j_a)$ et $(k1j_b)$. Soit $e_{hk} = 2(w_{h1j_a}y_{h1j_a} - w_{k1j_b}y_{k1j_b})$ la différence entre les contributions des deux USE à z_{hi} dans (11). Posons que $z_{h1(j_a)} = \sum_{j\neq j_a} 2w_{h1j}z_{h1j}$ et $z_{k1(j_b)} = \sum_{j\neq j_b} 2w_{k1j}z_{k1j}$ représentent, respectivement, z_{hi} en excluant les contributions des USE qui seront permutées. En notant que $z^*_{h1} = z_{h1} - e_{hk}$, $z^*_{h2} = z_{h2}$, $z^*_{k1} = z_{k1} + e_{hk}$, il découle de (12) que

$$\begin{aligned} 4[v(\hat{\bar{Y}} \mid S^*) - v(\hat{\bar{Y}} \mid S)] &= (z^*_{h1} - z^*_{h2})^2 - (z_{h1} - z_{h2})^2 \\ &\quad + (z^*_{k1} - z^*_{k2})^2 - (z_{k1} - z_{k2})^2 \\ &= 2e_{hk}[e_{hk} - (z_{h1} - z_{h2}) + (z_{k1} - z_{k2})] \end{aligned}$$

et, donc, l'équation (13) est vérifiée avec $g_{hk} = 2^{-1}\{[z_{h2} - z_{h1(j_a)}] - [z_{k2} - z_{k1(j_b)}]\}$. Les preuves pour les trois autres cas sont semblables. Quand $h = k$, nous avons $e_{hh} = 2(w_{h1j_a}z_{h1j_a} - w_{h1j_b}z_{h1j_b})$, $z^*_{h1} = z_{h1} - e_{hh}$ et $z^*_{h2} = z_{h2} - e_{hh}$. La preuve est achevée en posant $g_{hh} = 2^{-1}\{[z_{h2(j_b)} - z_{h1(j_a)}] - [z_{h1(j_a)} - z_{h2(j_b)}]\} = z_{h1(j_b)} - z_{h2(j_b)}$.

Preuve de l'équation (17)

Par définition, nous avons $z_{hi} = z_{1,hi} + z_{2,hi}$ pour tout h et i. Donc, en observant que $z^\dagger_{h1} - z^\dagger_{h2} = z_{h1} - z_{h2} + 2(z_{1,h2-z_{2,h1}})$ et $z_{h1} - z_{h2} + z_{1,h2} - z_{2,h1} = z_{1,h1} - z_{2,h2}$, nous avons

$$\begin{aligned} v(\hat{\bar{Y}} \mid S^\dagger) &= \sum_{h=1}^{H}\left(\frac{z^\dagger_{h1} - z^\dagger_{h2}}{2}\right)^2 \\ &= \sum_{h=1}^{H}\left[\left(\frac{z_{h1} - z_{h2}}{2}\right) + (z_{1,h2} - z_{2,h1})\right]^2 \\ &= \sum_{h=1}^{H}\left(\frac{z^\dagger_{h1} - z^\dagger_{h2}}{2}\right)^2 \\ &\quad + \sum_{h=1}^{H}(z_{1,h2} - z_{2,h1})(z_{h1} - z_{h2} + z_{1,h2} - z_{2,h1}) \\ &= v(\hat{\bar{Y}} \mid S) + \sum_{h=1}^{H}(z_{1,h1} - z_{2,h2})(z_{1,h2} - z_{2,h1}), \end{aligned}$$

ce qui complète la preuve.

Remerciements

Ces travaux ont été réalisés pendant que l'auteur était à Westat, Inc., aux États-Unis. L'auteur remercie Leyla Mohadjer, Sylvia Dohrmann, Jill Montaquila et Lexter R. Curtin de leur appui pour cette étude. Il remercie aussi Barry Graubard, le rédacteur associé, et deux examinateurs anonymes de leurs commentaires constructifs qui l'ont aidé à améliorer l'article.

Bibliographie

Dalenius, T., et Peiss, S.P. (1982). Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inferences*, 6, 73-85.

Dohrmann, S., Curtin, L.R., Mohadjer, L., Montaquila, J. et Le, T. (2002). National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 807-812.

Dohrmann, S., Lu, W., Park, I., Sitter, R. et Curtin, L.R. (2005). Variance estimation and data disclosure issues in the National Health and Nutrition Examination Survey limiting the risk of data disclosure using replication techniques in variance estimation. Soumis dans une revue.

Eltinge, J.L. (1999). Use of stratum mixing to reduce primary-unit-level identification risk in public-use survey datasets. *Proceedings of the 1999 Federal Committee on Statistical Methodology Research Conference*.

Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Gomatam, S., Karr, A.F. et Sanil, A.P. (2005). Data swapping as a decision problem. *Journal of Official Statistics*, 21, 635-655.

Lu, W. (2004). Confidentiality and Variance Estimation in Complex Surveys. Thèse de doctorat non publiée, Simon Fraser University, Department of Statistics and Actuarial Science.

Lu, W., Brick, M.J. et Sitter, R.R. (2006). Algorithms for constructing combined strata grouped jackknife and balanced repeated replications with domains. *Journal of the American Statistical Association*, 101, 1680-1692.

Mayda, J.E., Mohl, C. et Tambay, J.-L. (1996). Variance estimation and confidentiality: They are related! *Proceedings of the Survey Methods Section*, *Statistical Society of Canada*, 135-141.

National Center for Health Statistics (1994). Data file document, National Health Interview Survey of Topics Related to the Year 2000 Health Objectives, 1993 (Fichier de données exploitable à la machine et documentation), National Center for Health Statistics, Hyattsville, Maryland.

Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. et Curtin, L.R. (2006) Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1761-1767.

Park, I. (2004). Assessing complex sample designs via design effect decompositions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 4135-4142.

Park, I., et Lee, H. (2004). Effets de plan pour les estimateurs pondérés de la moyenne et du total sous échantillonnage complexe. *Techniques d'enquête*, 30, 205-216.

Rust, K.F. (1986). Efficient replicated variance estimation. In the *Proceedings of the Survey Research Methodos Section*, American Statistical Association, 81-87.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Shah, B. (2001). A Method to Create Pseudo Strata and PSU's based on BRR weights. Manuscript non-publié, Research Triangle Institute.

Shao, J., et Tu, D. (1995). *The Jackknife and Bootstrap*. New York : Springer-Verlag.

Valliant, R. (1996). Limitations of balanced half-sampling. *Journal of Official Statistics*, 12, 225-240.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Predication Approach*. New York : John Wiley & Sons, Inc.

Yung, W. (1997). Variance estimation for public use files under confidentiality constraints. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 434-439.

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 217-226
Statistique Canada, N° 12-001-X au catalogue

Une approche arborescente de la formation de strates dans les enquêtes-entreprises polyvalentes

Roberto Benedetti, Giuseppe Espa et Giovanni Lafratta [1]

Résumé

Pour concevoir un échantillon aléatoire simple stratifié sans remise à partir d'une population finie, il faut résoudre deux grandes questions : définir une règle de partition de la population en strates et répartir les unités d'échantillonnage entre les strates sélectionnées. Dans le présent article, nous examinons une stratégie arborescente en vue d'aborder conjointement ces deux questions quand l'enquête est polyvalente et que de l'information multivariée, quantitative ou qualitative, est disponible. Nous formons les strates à l'aide d'un algorithme divisif hiérarchique qui sélectionne des partitions de plus en plus fines en minimisant, à chaque étape, la répartition d'échantillon requise pour atteindre les niveaux de précision établis pour chaque variable étudiée. De cette façon, nous pouvons satisfaire un grand nombre de contraintes sans augmenter fortement la taille globale d'échantillon et sans écarter certaines variables sélectionnées pour la stratification ni diminuer le nombre de leurs intervalles de classe. En outre, l'algorithme a tendance à ne pas définir de strate vide ou presque vide, ce qui évite de devoir regrouper certaines strates. Nous avons appliqué la méthode au remaniement de l'Enquête sur la structure des exploitations agricoles en Italie. Les résultats indiquent que le gain d'efficacité réalisé en utilisant notre stratégie n'est pas trivial. Pour une taille d'échantillon donnée, cette méthode permet d'obtenir la précision requise en exploitant un nombre de strates qui est habituellement égal à une fraction très faible du nombre de strates disponibles quand on combine toutes les classes possibles provenant de n'importe quelle covariable.

Mots clés : Stratification multivariée ; répartition polyvalente optimale de l'échantillon ; enquête sur la structure des exploitations agricoles ; plan d'échantillonnage.

1. Introduction

De nombreuses enquêtes-entreprises s'appuient sur des méthodes d'échantillonnage stratifiées dans lesquelles un échantillonnage aléatoire simple sans remise est exécuté dans chaque strate (voir, par exemple, Sigman et Monsour 1995 et, pour les enquêtes sur les fermes, Vogel 1995). Habituellement, la base de sondage à partir de laquelle sont sélectionnées les unités est établie en utilisant des données administratives ou des données de recensement, représentées par une riche base de données sur des variables auxiliaires, qui sont chacune susceptibles d'être exploitées pour former les strates. De surcroît, les enquêtes de ce genre sont souvent polyvalentes et les estimations des multiples variables étudiées doivent atteindre des niveaux de précision donnés.

Satisfaire un aussi grand nombre de contraintes sans accroître considérablement la taille d'échantillon est un objectif généralement considéré comme strictement relié au choix du nombre de variables de stratification et de leurs intervalles de classe (Kish et Anderson 1978). Cette façon de voir tient au fait bien connu que des partitions plus fines de la population introduisent une plus grande quantité d'information utile pour réduire les variances des estimations ; par contre, leur application accroît le risque que les unités deviennent des « sauteuses » de strate.

Dénommons stratification *atomisée* celle obtenue en formant les strates par combinaison de toutes les classes possibles pour n'importe laquelle des covariables utilisées. Si le nombre correspondant de ces strates de base, ou atomes, excède un seuil donné imposé par des restrictions d'ordre pratique, il semble inévitable de devoir reconcevoir l'enquête en sélectionnant un nombre plus faible de variables de stratification ou en créant un moins grand nombre de classes pour chacune d'elles. Néanmoins, il convient de noter qu'un autre moyen d'obvier à cette situation indésirable est fondé sur l'argument suivant : nous pouvons vraiment interpréter la stratification atomisée comme une solution extrême du problème de formation des strates, puisque, entre le cas de l'absence de stratification et celui de l'utilisation de la stratification atomisée, il existe toute une gamme de possibilités de sélectionner une stratification dont les sous-populations peuvent être formées par des unions d'atomes.

Nous proposons, pour réaliser cette sélection, de définir un plan d'échantillonnage stratifié à structure arborescente. Nous formons les strates à l'aide d'un algorithme divisif hiérarchique qui sélectionne des partitions de plus en plus fines en minimisant, à chaque étape, la répartition d'échantillon requise pour atteindre les niveaux de précision fixés pour chaque variable étudiée. La procédure est séquentielle et détermine un chemin allant de la stratification nulle,

1. Roberto Benedetti, professeur, Département des sciences commerciales, statistiques, technologiques et environnementales (DASTA), Université « G. d'Annunzio », Viale Pindaro 42, Pescara, IT-65127. Courriel : benedett@unich.it ; Giuseppe Espa, professeur, Département d'économie, Université de Trente, Via Inama 5, Trente, IT-38100. Courriel : giuseppe.espa@economia.unitn.it ; Giovanni Lafratta, professeur adjoint, Département des sciences commerciales, statistiques, technologiques et environnementales (DASTA), Université « G. d'Annunzio », Viale Pindaro 42, Pescara, IT-65127. Courriel : giovanni.lafratta@unich.it.

c'est-à-dire celle dont la strate unique correspond à la population, à la stratification atomisée. À chaque étape, nous sélectionnons la variable qui doit être utilisée pour définir la nouvelle partition plus désagrégée : chaque strate de la partition courante est divisée sur une covariable, en utilisant tour à tour toutes les classes de cette dernière disponibles, et la covariable qui réduit le mieux la taille globale allouée est celle qui est choisie.

Bloch et Segal (1989) ont discuté de l'application des méthodes arborescentes de classification (voir, par exemple, Breiman, Friedman, Olshen et Stone 1984) à la formation des strates, mais ils s'intéressaient principalement à l'interprétation au niveau des strates des relations entre les covariables et une seule variable de résultat. Au contraire, nos règles de partition de la population sont axées directement sur la répartition optimale des unités d'échantillonnage dans les strates sélectionnées. Les méthodes classiques traitant le cas univarié (Dalenius et Hodges 1959 ; Singh 1971 ; Lavallée et Hidiroglou 1988 ; Hedlin 2000 ; Lu et Sitter 2002 ; Gunning et Horgan 2004 ; pour une revue voir Horgan 2006) ne peuvent pas être étendues facilement au cas où l'on cherche à exploiter plusieurs covariables pour la stratification. Par conséquent, les solutions proposées par ces auteurs ont peu d'utilité pratique si l'enquête est polyvalente et que l'on dispose d'information sur de multiples covariables. Dans un tel contexte, des méthodes visant à satisfaire un grand nombre de contraintes sur les erreurs lorsqu'on minimise la taille d'échantillon ont été proposées par Bethel (1985, 1989) et Chromy (1987). Valliant et Gentle (1997) ont également abordé le problème dans le cas de cadres d'échantillonnage à deux degrés. Pour une stratification donnée, nous choisissons d'appliquer la règle de répartition de Bethel et, par conséquent, la procédure consiste à sélectionner les partitions subséquentes en minimisant la fonction de coût de l'enquête correspondant aux stratifications constituées par les strates non partitionnées au moment de la minimisation et les sous-strates partitionnées disponibles.

Comme nous l'avons dit auparavant, dans le cadre général de la stratification multivariée, notre but est, à l'instar de Kish et Anderson (1978), d'utiliser certains résultats obtenus dans le domaine de l'échantillonnage stratifié en vue de répondre aux besoins des praticiens d'enquête. Ces derniers exécutent quotidiennement des enquêtes multivariées (plusieurs variables disponibles pour la stratification) et polyvalentes (plusieurs variables et de nombreuses autres statistiques sont les objectifs principaux des activités de l'enquête). Donc, l'objectif de notre approche est de permettre de combiner la stratification et la répartition de l'échantillon. Autrement dit, nous nous soucions du choix du nombre de variables de stratification, du nombre d'intervalles de classe pour chaque variable et de la répartition optimale de Bethel dans les strates. Étant donné ce choix, notre méthodologie ne peut être réduite à la solution classique du problème de stratification multivariée, c'est-à-dire le recours à des techniques multivariées, telles que l'analyse en grappes et l'analyse en composantes principales (voir, par exemple, Mulvey 1983, Pla 1991 et Jarque 1981). En fait, dans cette branche de la littérature, les auteurs n'utilisent pas (ou n'utilisent qu'indirectement) les variables d'intérêt et se servent uniquement des variables auxiliaires, et ils n'abordent pas la question de la répartition. Il serait encore plus difficile de justifier de réduire notre approche à celles passées en revue par Särndal, Swensson et Wretman (1992, sections 12.6 et 12.7) : les techniques présentées à la section 12.6 combinent la stratification et la répartition multivariée de l'échantillon, mais elles ne sont pas polyvalentes, tandis que les méthodes de la section 12.7 sont polyvalentes, mais sont basées sur des strates prédéterminées.

La présentation de l'article est la suivante. À la section 2, nous présentons la procédure que nous proposons pour le calcul des arbres de stratification. Nous décrivons en détail l'algorithme utilisé pour générer la suite de stratifications et nous montrons comment il peut être représenté sous forme d'un arbre de classification. Nous discutons aussi des critères d'arrêt, afin de déterminer leur incidence sur le nombre optimal de strates. À la section 3, nous examinons les moyens d'exploiter un arbre de stratification pour concevoir l'Enquête sur la structure des exploitations agricoles (ESEA) de l'Union européenne. Nous illustrons notre technique de stratification en produisant un ensemble arborescent de strates et de répartitions d'échantillon en nous servant d'un ensemble d'atomes de base définis au moyen de l'information multivariée recueillie durant le cinquième Recensement général de l'agriculture réalisé en Italie en l'an 2000. Enfin, à la section 4, nous présentons certaines conclusions, en nous concentrant sur les questions d'ordre pratique ayant trait à la formation de strates à l'aide d'arborescences et en exposant comment la procédure peut être utilisée pour mieux gérer les enquêtes polyvalentes basées sur des plans de sondage stratifiés.

2. Une méthode de création d'arbres de stratification multivariés

Considérons une population finie P de N unités sur lesquelles les variables $y_1, ..., y_g, ..., y_G$ doivent être étudiées pour estimer leurs totaux en se servant d'une stratification sur P, c'est-à-dire un ensemble F de sous-populations non vides H_F, appelées strates, qui partitionnent P. Notre problème est de savoir comment sélectionner F afin de minimiser la taille globale allouée de l'échantillon correspondant n_F de telle façon que, pour $g = 1, ..., G$, le coefficient de variation (CV_g) correspondant à la

g^{e} variable d'intérêt ne soit pas supérieur au niveau de précision souhaité, disons $\varepsilon_g > 0$.

Pour un F donné, cette minimisation est exécutée en calculant la règle de répartition de l'échantillon de Bethel (1985). Plus précisément, désignons par n_h, $h = 1, ..., H_F$, la part de l'échantillon dans la strate h. Le coût global d'enquête correspondant à F peut alors être donné par

$$f(n_1, ..., n_{H_F}) = c_F + \sum_{h=1}^{H_F} c_h n_h,$$

où c_F est un coût fixe indépendant de $\mathbf{n}_F = (n_1, ..., n_{H_F})'$, et c_h représente le coût d'échantillonnage d'une unité dans la strate h. De surcroît, soit Y_g le total dans P de la g^{e} variable de réponse, N_h la taille de la h^{e} strate de F et $S^2_{h,g}$ la variance de y_g dans la strate h. Alors, la g^{e} contrainte sur la précision requise peut s'exprimer :

$$(\text{CV}_g)^2 \le \varepsilon_g^2 \equiv \sum_{h=1}^{H_F} \frac{N_h^2 S^2_{h,g}}{n_h} - \sum_{h=1}^{H_F} N_h S^2_{h,g} \le Y_g^2 \varepsilon_g^2$$

$$\equiv \sum_{h=1}^{H_F} \frac{N_h^2 S^2_{h,g}}{\left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S^2_{h,g}\right) n_h} \le 1,$$

de sorte que, si nous considérons les quantités suivantes, que nous dénommons unités normalisées de précision,

$$\xi_{h,g} = N_h^2 S^2_{h,g} \Big/ \left(Y_g^2 \varepsilon_g^2 + \sum_{h=1}^{H_F} N_h S^2_{h,g}\right),$$

le problème de la répartition optimale de F peut s'exprimer de la façon suivante :

$$\begin{array}{lll} \min & f(\mathbf{n}_F) & \\ \text{sous les contraintes} & \sum_{h=1}^{H_F} \xi_{h,g} / n_h \le 1, & g = 1, ..., G, \\ & 1 / n_h > 0, & h = 1, ..., H_F. \end{array}$$

Bethel (1989) a obtenu la solution de ce genre de problème, disons n_h^*, $h = 1, ..., H_F$, comme il suit :

$$1 / n_h^* =$$

$$\begin{cases} \sqrt{c_h} \Big/ \left(\sqrt{\sum_{g=1}^{G} \alpha_g^* \xi_{h,g}} \sum_{l=1}^{H_F} \sqrt{c_l \sum_{g=1}^{G} \alpha_g^* \xi_{l,g}} \right) & \text{si } \sum_{g=1}^{G} \alpha_g^* \xi_{h,g} > 0, \\ +\infty & \text{autrement,} \end{cases}$$

où $\alpha_g^* = \lambda_g / \sum_{g=1}^{G} \lambda_g$, et λ_g est le multiplicateur de Lagrange de la contrainte sur l'erreur maximale permise dans l'estimation de la g^{e} variable étudiée, et indique si la g^{e} contrainte est « active » dans la solution du problème de répartition (à savoir, si $\alpha_g^* = 0$, alors la contrainte n'est pas active). La répartition optimale globale correspondante est donc donnée en fixant $n_F = \sum_{h=1}^{H_F} n_h^*$.

Supposons maintenant que les totaux estimés et leurs variances sont connus pour un ensemble donné de $M > 1$ strates de base $A_1, ..., A_m, ..., A_M$, de sorte que nous pouvons nous appuyer sur deux matrices $M \times G$ de totaux $\mathbf{T} = (Y_{m,g})$ et de variances des estimations $\mathbf{V} = (S^2_{m,g})$, respectivement, et sur les tailles N_m, $m = 1, ..., M$. La définition de ces strates, que nous nommerons dans la suite *atomes*, est basée sur un ensemble de covariables $X_1, ..., X_k, ..., X_K$ de la façon suivante. Soit $x_{i,k}$ la valeur de X_k mesurée sur l'unité $i \in P$, et considérons l'ensemble de valeurs observées distinctes pour X_k dans P, $\Xi_k = \{x \in \mathbb{R} : \exists i \in P : x = x_{i,k}\}$. Nous construisons $M = \prod_{k=1}^{K} |\Xi_k|$ atomes, un pour chaque vecteur $(a_{m,1}, ..., a_{m,K})$ dans le produit cartésien $\Xi = \otimes_{k=1}^{K} \Xi_k$ en fixant pour $m = 1, ..., M$

$$A_m = \bigcap_{k=1}^{K} A_{m,k},$$

où $A_{m,k} = \{i \in P : x_{i,k} = a_{m,k}\}$. Dans le cas où les covariables X_k sont continues, l'ensemble Ξ_k contiendra N atomes non vides. Comme l'algorithme est divisif hiérarchique, le nombre d'atomes finaux n'a aucune incidence sur les étapes de l'algorithme. Seule la phase initiale de construction de la statistique agrégée et, au plus, la répartition de la mémoire sont affectées. Notre expérience empirique donne à penser que le temps de calcul varie peu si la taille des atomes est égale à 1. Au contraire, les variables continues ou ordonnées accélèrent l'algorithme, car le nombre de partitions binaires possible est, si le nombre de valeurs est le même, beaucoup plus faible que dans le cas des variables catégoriques. Nous vérifions que cette construction produit une stratification : chaque unité de la population figure dans un atome, et un atome seulement. Afin d'illustrer nos définitions, référons-nous aux données présentées au tableau 1, où est décrit un simple exemple dans lequel il est supposé qu'un ensemble de $M = 9$ atomes (obtenus en exploitant $K = 2$ covariables possédant toutes deux trois valeurs distinctes, à savoir 1, 2 et 3) constitue la stratification de base pour étudier $G = 2$ variables, dont les totaux et les variances des estimations sont également donnés, ainsi que les tailles d'atome. Dans ce contexte, nous avons $\Xi_1 = \Xi_2 = \{1, 2, 3\}$ et, par exemple, A_8 est la sous-population dont les éléments i sont tels que $x_{i,1} = a_{8,1} \equiv 3$ et $x_{i,2} = a_{8,2} \equiv 2$.

Tableau 1
Exemple de données pour 9 atomes et 2 variables étudiés

Atomes				Variables étudiées			
Id	Définition		Tailles	Totaux		Variances	
m	$a_{m,1}$	$a_{m,2}$	N_m	$Y_{m,1}$	$Y_{m,2}$	$S^2_{m,1}$	$S^2_{m,2}$
1	1	1	1 000	10	10	16	25
2	1	2	1 000	10	10	16	4
3	1	3	1 000	10	10	16	4
4	2	1	1 000	10	10	16	25
5	2	2	1 000	10	10	16	4
6	2	3	1 000	10	10	16	4
7	3	1	1 000	10	10	4	25
8	3	2	1 000	10	10	4	16
9	3	3	1 000	10	10	4	16

La méthode que nous proposons produit une suite de stratifications qui peuvent être représentées comme un arbre de classification. Définissons le niveau l d'un nœud donné ν dans l'arbre comme étant le nombre d'arcs dans la chaîne (unique) raccordant le nœud ν au nœud racine, et désignons par r_l le nombre de nœuds partageant le même niveau l. Puisqu'un seul nœud sera divisé à chaque niveau, nous avons $r_l = l + 1$ pour chaque l. À chaque niveau $l \geq 0$, la méthode détermine une classe F_l de r_l sous-populations non vides en lesquelles P peut être partitionnée, en les plaçant dans une correspondance bijective avec les nœuds de niveau l. Les strates contenues dans F_l sont toutes des candidates à la partition sur toute covariable donnée X_k et, en suivant Bethel (1989), nous calculons la répartition d'échantillon qui minimise de manière optimale la fonction de coût d'enquête pour la stratification constituée des strates non divisées dans F_l et des deux sous-strates qui définissent la partition en question. La meilleure partition au niveau l est celle qui est la plus favorable en ce sens qu'elle réduit l'échantillon alloué, par rapport à celui caractérisant F_l, plus que toute autre partition possible sur n'importe laquelle des covariables utilisées. La répartition optimale correspondant à la stratification définie par cette partition la meilleure, indiquée par $n_{b,\,l+1}$, est adoptée comme taille d'échantillon optimale au niveau $l + 1$ et considérée comme une borne supérieure contraignant les répartitions au niveau suivant de classification. À l'étape de l'initialisation, nous fixons $F_0 = \{P\}$, dont la strate unique est donc équivalente à l'ensemble de la population, et la meilleure taille d'échantillon $n_{b,0}$ est calculée comme étant le maximum parmi les tailles optimales obtenues en tenant compte, séparément, de chaque niveau de précision individuel ε_g fixé au sujet de la g^{e} variable étudiée :

$$n_{b,0} = \max_{g = 1, \ldots, G} \frac{N^2 S_g^2}{Y_g^2 \varepsilon_g^2 + N S_g^2},$$

où Y_g est l'estimation du total pour y_g sur P et S_g^2 est la variance correspondante (voir, pour la répartition optimale avec un seul item, Cochran 1977, pages 97-106, et Särndal et coll. 1992, pages 104-109).

Quand $l > 0$, l'ensemble de strates F_{l-1} optimal à l'étape $l - 1$ est celui qui est analysé. La meilleure répartition d'échantillon à l'étape l, $n_{b,l}$ est fixée au départ égale à $n_{b,\,l-1}$ et, pour chaque strate $U \in F_{l-1}$ et chaque variable auxiliaire X_k, l'algorithme qui suit est exécuté. Soit A_U l'ensemble d'atomes contenus dans la strate courante U, de sorte que l'égalité $U = \bigcup A_U$ est vérifiée et soit $m(A)$ une fonction donnant l'indice attribué à tout atome $A (m(A) = m_0$ si, et uniquement si $A = A_{m_0})$; nous pouvons alors exprimer l'ensemble de valeurs pris par X_k pour les unités contenues dans tout atome de A_U comme il suit :

$$Q_k = \{q \in \mathbb{R} \colon \exists A \in A_U \colon\ q = a_{m(A),\,k}\}.$$

Si X_k est une variable ordonnée, pour chaque q dans Q_k autre que $\max(Q_k)$, la strate U est partitionnée en deux ensembles $U_1 = U_{q,1}$ et $U_2 = U_{q,2}$ de la façon suivante :

$$U_{q,1} = \bigcup \{A \in A_U \colon\ a_{m(A),\,k} \leq q\}$$

et $U_{q,2}$ est le complément relatif de $U_{q,1}$ dans U, c'est-à-dire l'ensemble de tous les $i \in U$ qui ne sont pas compris dans $U_{q,1}$:

$$U_{q,2} = U \setminus U_{q,1}.$$

Dans notre exemple, pour une strate U définie comme étant $A_1 \cup A_2 \cup A_8$, nous avons $A_U = \{A_1, A_2, A_8\}$, $Q_1 = \{1, 3\}$ et $Q_2 = \{1, 2\}$ (voir le tableau 1), de sorte que notre algorithme essaiera de diviser U en $U_1 = A_1 \cup A_2$ et $U_2 = A_8$ en utilisant X_1, et en $U_1 = A_1$ et $U_2 = A_2 \cup A_8$ en utilisant X_2. Si, au contraire, X_k n'est pas ordonnée, U est alors partitionné en les ensembles U_1 et U_2 pour chaque sous-ensemble approprié U_1 de U, avec $U_2 = U \setminus U_1$.

Nous avons donc une stratification candidate correspondante, à savoir

$$C = (F_{l-1} \setminus \{U\}) \cup \{U_1\} \cup \{U_2\},$$

qui inclut toutes les strates de F_{l-1} autres que U, et, en plus, U_1 et U_2. Pour chaque strate C dans l'ensemble C, les totaux estimés de Y_g, $g = 1, \ldots, G$,

$$Y_{C,g} = \sum_{A \in A_C} Y_{m(A),\,g},$$

et leurs variances correspondantes

$$S_{C,g}^2 = (N_C - 1)^{-1} \Big(\sum_{A \in A_C} (N_A - 1) S_{m(A),\,g}^2 + \sum_{A \in A_C} N_A (N_A^{-1} Y_{m(A),\,g} - N_C^{-1} Y_{C,g})^2 \Big),$$

sont calculés et la répartition d'échantillon n_C est donc obtenue en appliquant la règle de Bethel. Si $n_C < n_{b,l}$, alors la partition (U_1, U_2) devient la partition courante la meilleure, la meilleure stratification candidate C^* devient C et $n_{b,l}$ est mise à jour pour devenir n_C. De cette façon, la méthode divisive qui produit le meilleur résultat, c'est-à-dire la plus petite taille d'échantillon, est sélectionnée pour créer les strates optimales suivantes :

$$F_l = C^*.$$

Dans le cadre de notre exemple, pour les niveaux de précision $\varepsilon_1 = \varepsilon_2 = 0{,}1$, décrivons la partition optimale au

niveau $l = 1$, c'est-à-dire celle qui divise la population entière en deux strates. Si nous utilisons les données décrites au tableau 1, l'algorithme indique que la meilleure partition de $U = P$ est basée sur la variable X_2 et est obtenue en fixant

$$U_1 = \bigcup \{A \in A_P : a_{m(A),k} \leq 2\}$$

$$= A_1 \cup A_2 \cup A_4 \cup A_5 \cup A_7 \cup A_8$$

et, en conséquence, $U_2 = P \setminus U_1 = A_3 \cup A_6 \cup A_9$. Cette partition optimale est représentée à la figure 1, où sont indiquées en détail, pour chaque strate, la taille, la définition en termes d'atomes inclus, la répartition courante et les statistiques d'estimation.

Les questions concernant le nombre optimal de strates sont prises en compte en définissant le critères d'arrêt de la procédure de création de l'arborescence. Nous décidons d'arrêter l'algorithme si la différence relative entre la taille optimale d'échantillon au niveau courant et celle au niveau précédent est plus faible qu'un paramètre donné $\delta > 0$:

$$\delta > (n_{b,l-1} - n_{b,l}) / n_{b,l-1}. \qquad (1)$$

Comme l'algorithme de Bethel converge vers un vecteur dont l'étendue est $(0, +\infty)^{l+1}$, ses données d'entrée doivent être arrondies aux nombres entiers correspondants les plus proches vers l'infini ; par conséquent, surtout en présence d'un grand nombre de petites strates, une répartition donnée produira vraisemblablement une taille d'échantillon plus grande que la précédente. En outre, dans ce cas, nous avons décidé d'arrêter notre procédure. Pour éviter d'avoir des strates trop petites et, par conséquent, statistiquement instables, des règles supplémentaires peuvent être établies pour empêcher des désagrégations plus poussées des strates courantes si les sous-strates correspondantes ont une cardinalité plus faible qu'une taille de strate minimale prédéfinie. Il est également possible d'atténuer les complexités de la gestion de l'enquête en imposant un nombre maximal de strates.

Cette approche, qui consiste à exécuter une recherche exhaustive dans chaque partition, garantit que la stratification et la répartition correspondantes sont optimales, mais seulement conditionnellement aux partitions exécutées antérieurement. Nous savons que la monotonicité des solutions et l'optimalité conditionnelle de chaque sous-arborescence obtenue par partition récursive de chaque nœud sont des conditions nécessaires, mais non suffisantes pour qu'un arbre binaire soit optimal. Afin de s'assurer de l'optimalité globale, nous devons ajouter à ces conditions l'exigence qu'une stratification optimale en, disons, H strates ne puisse être obtenue que par partition d'un des nœuds de la stratification optimale en $H - 1$ strates. Autrement dit, nous devons supposer qu'une partition optimale en H strates est un sous-espace de la partition optimale en $H - 1$ strates, ce qui implique que partitionner une strate donnée ne modifiera pas la fonction objectif – c'est-à-dire la répartition – dans les $H - 1$ strates restantes. Cependant, cette hypothèse est rarement vérifiée dans les applications pratiques d'enquête, puisque la partition d'une strate induit habituellement une modification des répartitions optimales dans toutes les strates restantes non partitionnées.

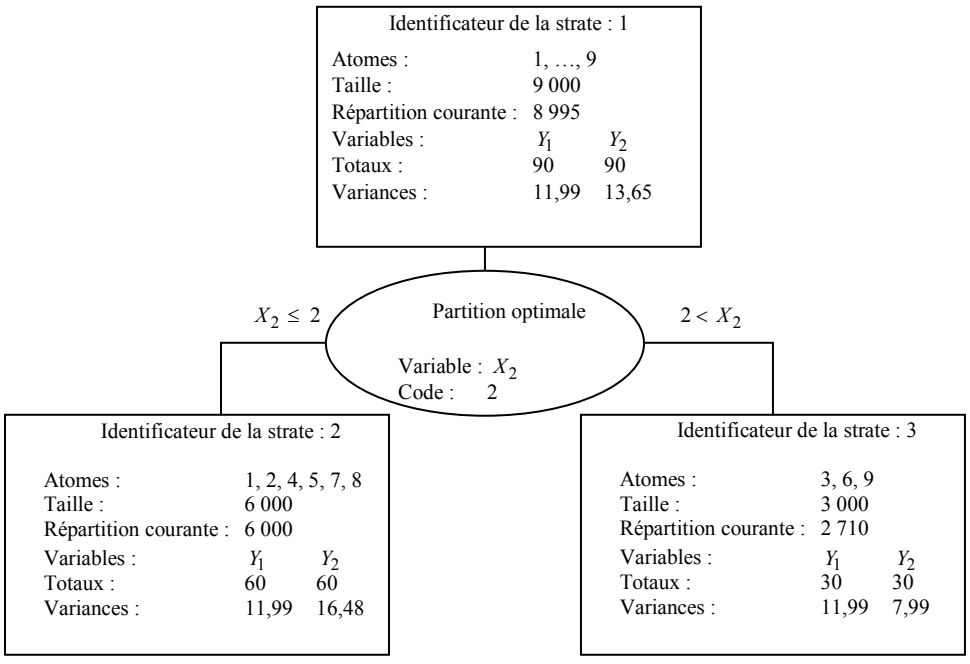


Figure 1 La première partition optimale pour les données utilisées comme exemple

L'algorithme que nous proposons, inspiré par la nature séquentielle et récursive des arbres binaires, peut être considéré comme une approche heuristique du problème de la stratification multivariée qui nous permet de déceler de bonnes strates, presque optimales, au prix d'un fardeau de calcul raisonnable. Par conséquent, cette méthode est efficace pour partitionner des populations en utilisant de grands ensembles de variables auxiliaires de stratification continues et qualitatives. En outre, la structure simple des arbres binaires offre une grande souplesse en ce qui concerne l'introduction de tout nombre de contraintes supplémentaires, telles que des bornes inférieures sur le nombre d'unités dans chaque strate.

3. Formation de strate pour l'Enquête sur la structure des exploitations agricoles en Italie

Conformément aux exigences des politiques agricoles de l'Union européenne, l'Enquête sur la structure des exploitations agricoles (ESEA) est exécutée tous les deux ans (règlement du Conseil (CEE) n° 70/66) en vue de recueillir des données sur des variables technoéconomiques caractéristiques des entreprises agricoles de l'UE. Elle représente la source principale d'information pour le projet EUROFARM (règlement du Conseil (CEE) n° 571/88), un ensemble de banques de données à utiliser pour traiter les enquêtes de l'UE sur la structure des exploitations agricoles. Les États membres sont tenus de prendre toutes les mesures appropriées pour exécuter l'ESEA sur leur territoire et sont libres de choisir un critère d'échantillonnage, mais le questionnaire et la précision requise, au niveau national, pour les estimations des variables étudiées sont déterminés par les règlements de la Communauté (voir les règlements CE n° 837/90 et n° 959/93, et les décisions 1998/377/CE et 2000/115/CE subséquentes de la Commission).

Afin d'illustrer notre méthode de stratification, nous exécutons l'algorithme décrit à la section 2 pour établir le plan de sondage de l'ESEA italienne et définir un ensemble arborescent de strates et de répartitions à l'aide d'information multivariée. Tous les algorithmes ont été implémentés par l'un des auteurs en langage MATLAB ; une application console Win32 a également été développée en C++ pour permettre des traitements par lot. Le plan de sondage exploite la liste des exploitations agricoles établies durant le cinquième Recensement général de l'agriculture tenu en Italie à l'automne 2000. ISTAT, l'Institut national de statistique de l'Italie, est responsable des mises à jour de cette liste par intégration des enregistrements administratifs, mais ces mises à jour n'étaient pas disponibles au moment de la rédaction de l'article. Pour initialiser la procédure, nous avons besoin d'un ensemble d'atomes dans lesquels la population d'exploitations agricoles italiennes doit être partitionnée. Nous obtenons cet ensemble de strates de base par agrégation des exploitations agricoles ayant en commun les mêmes classes pour sept covariables. Nous choisissons quatre variables associées à l'utilisation des terres et aux cheptels, à savoir la superficie agricole utile (SAU), le nombre de bovins (NB), le nombre de porcins (NP) et le nombre d'ovins et de caprins (NOC). Afin de tenir compte des caractéristiques géographiques des exploitations agricoles, nous avons également ajouté, comme variable de stratification, l'altitude de l'exploitation (ALT). Enfin, nous avons recueilli des renseignements au sujet de l'administration et de l'organisation de l'exploitation au moyen de deux variables nommées personnalité juridique de l'exploitant (PJ) et mode de tenure de l'exploitation (MT).

Les étendues des covariables concernant la structure de l'exploitation agricole sont réparties en quatre classes pour le nombre de bovins ($NB = 0$, $1 \le NB < 10$, $10 \le NB < 50$, $50 \le NB$), le nombre de porcins ($NP = 0$, $1 \le NP < 500$, $500 \le NP < 1{,}000$, $1{,}000 \le NP$) et le nombre d'ovins et de caprins ($NOC = 0$, $1 \le NOC < 250$, $250 \le NOC < 500$, $500 \le NOC$) et en sept classes pour la superficie agricole utile ($SAU = 0$, $0 < SAU < 1$, $1 \le SAU < 5$, $5 \le SAU < 10$, $10 \le SAU < 50$, $50 \le SAU < 100$, $SAU \ge 100$ ha). L'étendue des valeurs de l'altitude est divisée en cinq classes : montagnes intérieures, montagnes côtières, collines intérieures, collines côtières et plaines. Les classes pour la personnalité juridique de l'exploitant sont définies de manière à faire la distinction entre les titulaires exclusifs, les personnes morales (sociétés) et les groupes de personnes physiques (partenariat) dans un groupe, les entreprises coopératives, les associations d'exploitants, les institutions publiques et, enfin, les personnalités juridiques autres que celles qui précèdent (par exemple, consortia), que nous nommeront personnalités juridiques résiduelles. Les exploitations agricoles sont également stratifiées en tenant compte du mode de tenure, en faisant la distinction entre les propriétaires-exploitants (avec des sous-classes supplémentaires basées sur les catégories de travailleurs agricoles : membres de la famille, principalement membres de la famille, principalement non-membres de la famille), les locataires-exploitants, le partage des superficies agricoles exploitées et les modes de tenure autres que ceux qui précèdent. La combinaison de toutes les classes possibles pour chacune des covariables sélectionnées aboutit à 2 964 atomes non vides, qui constituent le point de départ de la procédure.

Nous avons étudié 12 variables d'utilisation des terres, dont la liste est présentée au tableau 2. Pour chaque variable étudiée, nous avons calculé les totaux et les variances dans chaque atome à l'aide des données de recensement disponibles, ce qui nous a permis d'exécuter l'algorithme de Bethel à chaque étape de notre procédure. Les paramètres supplémentaires requis pour définir nos critères d'arrêt sont

établis comme il suit. Le nombre maximal de strates est fixé à 300 et nous avons décidé d'interdire les strates dont la taille est inférieure à 10. Nous introduisons une tolérance quant à l'écart relatif entre les tailles d'échantillon optimales aux niveaux subséquents en fixant $\delta = 0$ dans l'équation (1), de sorte que l'algorithme s'arrête si $n_{b,l-1} < n_{b,l}$ pour un niveau donné $l \geq 0$.

Tableau 2
Variables étudiées dans l'Enquête sur la structure des exploitations agricoles en Italie et leurs niveaux de précision

Variable étudiée	CV requis		
	Requis par l'ESEA	Réalisé par	
		stratification atomisée	arbre de stratification
Céréales	1,00	0,98	0,98
Vignes	3,00	1,38	1,38
Oliviers	3,00	1,11	1,11
Racines fourragères et herbacées	3,00	2,39	2,40
Plantes industrielles	3,00	2,22	2,23
Plantes fourragères	3,00	1,37	1,39
Légumes	3,00	3,03	3,03
Jachères	3,00	2,69	2,78
Nombre de bovins	1,00	0,99	1,00
Nombre de porcins	2,00	0,80	0,82
Nombre d'ovins	2,00	1,99	2,01
Nombre de caprins	2,00	1,92	1,98

Nous avons atteint la convergence puisque le nombre maximal de strates a été obtenu et qu'aucune règle d'arrêt n'a été activée pour $l < 300$. La figure 2 représente les répartitions optimales $n_{b,l}$ en fonction du nombre de strates r_l, $l = 1, ..., 300$ sur une échelle logarithmique, c'est-à-dire en fonction de $\log(n_{b,l})$. Il convient de souligner que l'écart relatif entre les répartitions subséquentes diminue rapidement, les dix premières partitions étant les plus importantes en ce qui concerne ce comportement : en fait, en fixant $\delta = 10\,\%$, la procédure atteindrait la convergence à l'étape $l = 7$.

La figure 3 donne un diagramme de l'arbre de stratification créé jusqu'au niveau 7. Afin d'optimiser la répartition globale, notre critère de partition a créé récursivement des strates de plus en plus petites. La première partition est faite sur la personnalité juridique de l'exploitant, PJ, et les atomes ont été inclus dans la strate fille de gauche si la classe de la variable PJ qu'ils supposaient était titulaire exclusif, institution publique ou PJ résiduelle. Cette partition est celle qui est optimale au niveau 1, puisqu'elle correspond à une partition de la population entière, la seule strate disponible au niveau 0, qui décrit le mieux la répartition de l'échantillon. Cette partition indique seulement que les exploitations agricoles dont l'exploitant est le titulaire exclusif se comportent différemment de celles gérées par des personnalités juridiques plus complexes, comme des sociétés, des partenariats, des associations ou des entreprises coopératives. La deuxième partition est faite sur le nombre de bovins, NB. Elle crée de nouvelles sous-strates de la strate 2 (voir la partie inférieure de la figure 3), à savoir les strates 4 et 5, de la façon suivante : la nouvelle strate 4 est définie comme l'union des atomes de la strate 2 pour lesquels la condition $NB > 10$ est vérifiée, tandis que la strate 5 est le complément relatif de la strate 4 dans la strate 2. De cette façon, l'algorithme détecte la meilleure diminution de la taille globale de l'échantillon (passant de 1 570 313 à 689 404 unités échantillonnées, voir la partie droite de la figure 3) en reconnaissant que les exploitations agricoles caractérisées par un cheptel bovin de taille moyenne ou grande doivent être traitées séparément pour celles dont l'exploitant est titulaire exclusif. La troisième partition est faite sur la superficie agricole utile, SAU. Ici, la strate 4 est partitionnée en une strate contenant les atomes pour lesquels la variable SAU est inférieure à 100 ha (strate 6) et une strate contenant les atomes restants (strate 7). Ces deux nouvelles strates sont à leur tour divisées, par étape successive, c'est-à-dire les étapes 4 à 7 (voir la partie gauche de la figure 3) sur les variables NP et NOC : plus précisément, la procédure suggère de faire la distinction entre les entreprises agricoles ne possédant pas de cheptel ovin et caprin ($NOC = 0$), ou caractérisées par un grand cheptel porcin ($NP \geq 500$).

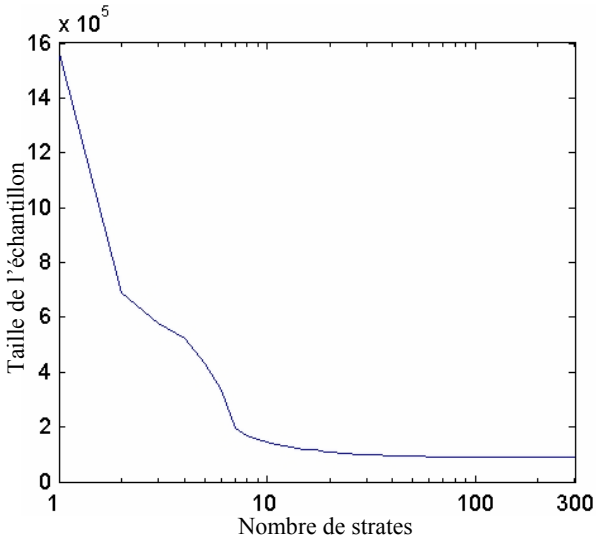


Figure 2 Tailles d'échantillon étape par étape. Les répartitions optimales $n_{b,l}$ sont montrées en fonction du nombre de strates r_l exploitées par le plan d'échantillonnage à structure arborescente aux étapes $l = 0, \ldots, 299$. Une échelle logarithmique est appliquée à l'axe horizontal, de sorte que $n_{b,l}$ est tracé en fonction de $\log(r_l)$. À mesure que le nombre de strates augmente, le plan de stratification à structure arborescente atteint ses objectifs en utilisant une taille globale d'échantillon rapidement décroissante, puisque la procédure améliore considérablement l'efficacité d'échantillonnage au cours de ses dix premières étapes d'exécution

Pour évaluer l'efficacité du plan d'échantillonnage à structure arborescente, nous calculons la meilleure répartition correspondant à la stratification atomisée, qui a déterminé un échantillon de 89 522 unités. En inspectant l'arbre de stratification, nous pouvons constater qu'une répartition globale fort semblable correspond à la meilleure stratification obtenue au niveau $l = 102$: en fait, pour cette partition de 103 strates, la taille d'échantillon est égale à 89 509. Cela signifie que, pour la même taille d'échantillon, notre algorithme atteint la précision requise pour l'enquête en exploitant un nombre de strates, 103, qui ne représente qu'une très petite fraction de 2 964, le nombre d'atomes disponibles, ce qui facilite l'organisation de l'enquête. Un autre avantage appréciable de notre procédure consiste à éviter les strates instables : il mérite d'être souligné que 1 618 des 2 964 atomes ont une taille égale ou inférieure à 5, alors que la taille minimale de n'importe laquelle des strates optimales au niveau 102 est de 16, si bien qu'il n'est pas nécessaire d'introduire des procédures de fusion de strates. Nous pouvons pousser plus loin les comparaisons et examiner les niveaux de précisions atteints en exécutant, respectivement, la stratification atomisée et l'arbre de stratification à l'étape 102. Ces niveaux, qui sont donnés au tableau 2, peuvent être considérés comme fort semblables pour les deux plans d'échantillonnage. En fait, nous constatons que, pour la stratification atomisée, la répartition de Bethel était activement contrainte sur la précision concernant trois variables étudiées, à savoir les céréales, les légumes et le nombre d'ovins. En ce qui concerne les strates correspondant au niveau 102 de l'arbre, les contraintes antérieures se trouvent aussi être actives, même si une autre contrainte, celle sur la variable Nombre de caprins a aussi donné des résultats serrés pour l'optimisation, les niveaux de précision atteints étant passés de 1,92 % à 1,98 %. Ces résultats donnent à penser qu'en ce qui concerne la partition atomisée, l'arbre peut être utilisé pour déceler une stratification plus compacte de la population, tout en préservant les niveaux de précision atteints et la taille globale d'échantillon.

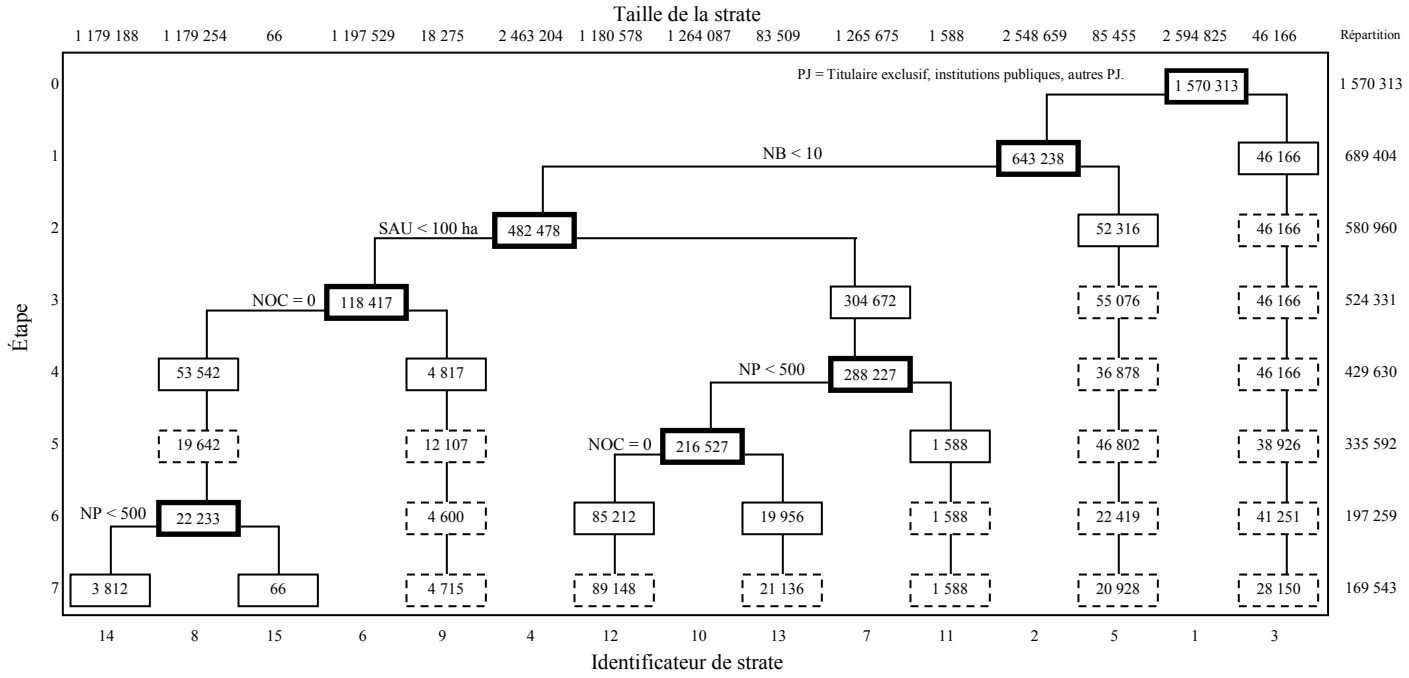


Figure 3 Diagramme de l'arbre de stratification. À la partie inférieure, l'axe horizontal représente l'identificateur de strate, un nombre qui indique de manière unique la sous-population correspondante dans la procédure de stratification. Les tailles de ces strates sont données à la partie supérieure de la figure. À la partie gauche, l'axe vertical représente la suite d'étapes allant de 0 à 7, tandis qu'à la partie droite, il représente la répartition globale optimale correspondant à ces étapes. Les blocs à bordure double représentent des strates partitionnées. Les strates filles sont reliées à leurs parents par des lignes coudées, et, quand elles ne sont plus partitionnées davantage aux étapes subséquentes, elles sont représentées comme des blocs à bordure simple. Pour les strates filles de gauche, la covariable sur laquelle la partition a été faite et la condition qu'elle a satisfait en définissant les sous-strates de gauche sont indiquées au-dessus de la ligne coudée correspondante. Le nombre à l'intérieur d'un bloc donné est la part de l'échantillon que la procédure attribue à la strate correspondante durant l'étape à laquelle le bloc est positionné. Puisqu'une strate peut demeurer non partitionnée durant des étapes qui suivent celle à laquelle elle a été créée, que sa répartition d'échantillon peut varier d'une étape à l'autre, des blocs en traits interrompus sont utilisés pour indiquer les modifications des tailles d'échantillon de strate

4. Conclusion

La stratégie arborescente pour les enquêtes polyvalentes examinées dans le présent article est planifiée de manière à définir conjointement une règle pour partitionner la population et pour répartir les unités d'échantillonnage entre les strates formées en exploitant l'information multivariée, quantitative ou qualitative. Un algorithme divisif hiérarchique sélectionne des partitions plus fines en minimisant, à chaque étape, la répartition de l'échantillon nécessaire pour atteindre les niveaux de précision requis. De cette façon, de grands nombres de contraintes peuvent être satisfaites sans accroître considérablement le nombre de strates. En outre, les variables choisies pour la stratification ne sont pas écartées simplement sur la base de considérations pratiques et le nombre de leurs intervalles de classe n'est pas diminué. Qui plus est, l'algorithme évite de créer des strates vides ou presque vides, si bien qu'il n'est pas nécessaire de procéder à des agrégations après la stratification en vue de mieux évaluer les variances d'estimation dans les strates.

Néanmoins, notre proposition peut susciter certaines critiques. Théoriquement, notre méthode ne peut pas être considérée comme une généralisation multiréponse de la méthode bien connue de l'arbre de régression et de classification, dont le but est d'exploiter la relation entre les covariables et une variable de résultat unique. En fait, même si nous traitons des enquêtes polyvalentes, notre approche consiste à partitionner l'information disponible de façon à optimiser uniquement une variable, à savoir la répartition d'échantillon dans les strates. De plus, la stratégie d'échantillonnage obtenue par notre méthode ne représente pas nécessairement un optimum global : en fait, la procédure constitue un algorithme de sélection ascendante de strates et, par conséquent, la recherche de l'optimalité à une étape donnée est conditionnée sur la stratification en cours d'utilisation, c'est-à-dire celle basée sur les partitions exécutées antérieurement. Il n'existe aucune garantie que la stratification sélectionnée par la procédure à une certaine étape l sera celle qui est optimale, même uniquement parmi toutes les partitions possibles dans les $l+1$ sous-ensembles de la population. Dans certaines situations, d'autres méthodes, comme la programmation dynamique, peuvent être utilisées pour exécuter une recherche exhaustive efficace de la stratification globalement optimale (voir Bühler et Deutler 1975, et Lavallée 1988).

Nous avons appliqué notre méthode au remaniement de l'Enquête sur la structure des exploitations agricoles en Italie. Les résultats indiquent que notre stratégie donne lieu à des gains d'efficacité : pour une taille donnée d'échantillon, notre méthode permet d'atteindre la précision requise en exploitant un nombre de strates qui est habituellement égal à une très petite fraction du nombre de strates disponibles quand sont combinées toutes les classes possibles pour n'importe laquelle des covariables. En outre, si l'on permet de grands nombres de strates, l'algorithme détecte d'autres stratégies d'échantillonnage pour lesquelles les contraintes sont satisfaites avec des tailles d'échantillon plus faibles que celle qui correspond à la stratification atomisée. Le choix de l'échantillonnage final dépend manifestement de la fonction de coût global de l'enquête. À cette fin, il est possible de recourir aux arbres de stratification pour tenir compte du fait qu'un nombre croissant de strates implique habituellement un accroissement des coûts dû à des problèmes d'organisation de l'enquête, mais qu'il correspond aussi à des tailles d'échantillon plus petites, qui donnent lieu à une réduction des coûts unitaires. La formation des strates par une méthode arborescente peut donc faciliter la gestion de l'enquête, en tant qu'outil d'aide à la sélection du plan d'échantillonnage stratifié qui convient le mieux à la collecte d'information au sujet du phénomène multivarié étudié.

Remerciements

Les auteurs remercient un rédacteur associé et deux examinateurs de leurs commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

Bibliographie

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. Dans *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 209-212.

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 49-60.

Bloch, D.A., et Segal, M.R. (1989). Empirical comparison of approaches to forming strata – Using classification trees to adjust for covariates. *Journal of the American Statistical Association*, 84, 408, 897-905.

Breiman, L., Friedman, J.H., Olshen, R.A. et Stone, C.J. (1984). *Classification and Regression Trees*, Belmont, CA : Wadsworth International Group.

Bühler, W., et Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Chromy, J. (1987). Design optimization with multiple objectives. Dans *Proceedings of the Surveys Research Methods Section*, American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. New York : John Wiley & Sons, Inc.

Dalenius, T., et Hodges, J.L. Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*. 54, 285, 88-101.

Gunning, P., et Horgan, J.M. (2004). Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques. *Techniques d'enquête*, 30, 2, 177-185.

Hedlin, D. (2000). A procedure for stratification by the extended ekman rule. *Journal of Official Statistics*, 16, 1, 15-29.

Horgan, J.M. (2006). Stratification of skewed populations: A review. *Revue Internationale de Statistique*. 74, 1, 67–76.

Jarque, C.M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *Applied Statistics*. 30, 2, 163-169.

Kish, L., et Anderson, D.W. (1978). Multivariate and multipurpose stratification. *Journal of the American Statistical Association*, 73, 361, 24-34.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Virginie, août 1988, 646-651.

Lavallée, P., et Hidiroglou, M.A. (1988). Sur la stratification de population asymétriques. *Techniques d'enquête*, 14, 35-45.

Lu, W., et Sitter, R.R. (2002). Méthode pratique de stratification multiple par programmation linéaire. *Techniques d'enquête*, 28, 2, 215-224.

Mulvey, J.M. (1983). Multivariate stratified sampling by optimization. *Management Science*, 29, 6, 715-724.

Pla, L. (1991). Determining stratum boundaries with multivariate real data. *Biometrics*, 47, 4, 1409-1422.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer Verlag.

Sigman, R.S., et Monsour, N.J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods*, (Éds. B.G. Cox, D.A. Binder, B. Nanjamma Chinnappa, A. Christianson, M.J. Colledge et P.S. Kott). New York : John Wiley & Sons, Inc., 133-152.

Singh, R. (1971). Approximately optimum stratification on the auxiliary variable. *Journal of the American Statistical Association*, 66, 336, 829-833.

Valliant, R., et Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

Vogel, F.A. (1995). The evolution and development of agricultural statistics at the United States department of agriculture. *Journal of Official Statistics*, 11, 2, 161-180.

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 227-236
Statistique Canada, N° 12-001-X au catalogue

Détermination des bornes optimales de strate au moyen de la programmation dynamique

Mohammad G. Mostafa Khan, Niraj Nand et Nesar Ahmad [1]

Résumé

La stratification optimale est la méthode qui consiste à choisir les meilleures bornes qui rendent les strates intérieurement homogènes, étant donné la répartition de l'échantillon. Afin de rendre les strates intérieurement homogènes, celles-ci doivent être construites de façon que les variances de strate de la caractéristique étudiée soient aussi faibles que possible. Un moyen efficace d'y arriver, si l'on connaît la distribution de la principale variable étudiée, consiste à créer des strates en découpant l'étendue de la distribution à des points appropriés. Si la distribution des fréquences de la variable étudiée est inconnue, on peut l'approximer en se fondant sur l'expérience passée ou sur certains renseignements a priori obtenus au cours d'une étude récente. Dans le présent article, le problème de la détermination des bornes optimales de strate (BOS) est considéré comme étant le problème de la détermination des largeurs optimales de strate (LOS). Il est formulé comme un problème de programmation mathématique (PPM) consistant à minimiser la variance du paramètre de population estimé sous la répartition de Neyman en imposant que la somme des largeurs des strates soit égale à l'étendue totale de la distribution. La variable étudiée est considérée comme suivant une loi continue dont la densité de probabilité est triangulaire ou normale standard. Les PPM formulés, qui s'avèrent être des problèmes de décision à plusieurs degrés, peuvent alors être résolus en utilisant la méthode de programmation dynamique proposée par Bühler et Deutler (1975). Des exemples numériques sont présentés pour illustrer les calculs. Les résultats obtenus sont également comparés à ceux donnés par la méthode de Dalenius et Hodges (1959) dans le cas d'une distribution normale.

Mots clés : Échantillonnage aléatoire stratifié ; stratification optimale ; loi triangulaire ; loi normale standard ; problème de programmation mathématique ; problème de décision à plusieurs degrés ; méthode de programmation dynamique.

1. Introduction

L'aspect fondamental à prendre en considération dans la détermination des bornes optimales de strate (BOS) est que l'homogénéité interne des strates doit être aussi grande que possible, c'est-à-dire que les variances de strate σ_h^2 doivent être aussi faibles que possible, étant donné la répartition de l'échantillon. Quand on étudie une seule caractéristique et que l'on connaît la distribution de la variable étudiée, les BOS peuvent être déterminées en découpant l'étendue de cette distribution en des points appropriés. Ce problème de détermination des BOS a été discuté pour la première fois par Dalenius (1950), dans le cas où la variable étudiée proprement dite est utilisée comme variable de stratification. Il a présenté un ensemble d'équations minimales pouvant être résolues pour trouver les BOS. Malheureusement, en général, ces équations ne pouvaient pas être résolues, à cause de leur nature implicite. Par conséquent, plusieurs auteurs ont tenté d'obtenir les bornes approximatives de strate en utilisant des méthodes classiques. Le nombre de strates étant donné, Dalenius et Gurney (1951) ont proposé de déterminer les bornes des strates quand $W_h\sigma_h$ demeure constant, où W_h est le poids de la strate h. Mahalanobis (1952), ainsi que Hansen et Hurwitz (1953) ont proposé que l'on détermine les bornes des strates quand $W_h\mu_h$ demeure constant. Aoyama (1954) a suggéré d'appliquer une règle approximative et recommandé de produire des strates de même largeur $x_h - x_{h-1}$, où x_{h-1} et x_h sont les bornes de la strate h. Ekman (1959) a déterminé les bornes des strates sous la contrainte que $W_h(x_h - x_{h-1})$ soit constant. Dalenius et Hodges (1959) ont recommandé de construire les strates en prenant des intervalles égaux sur la cumulative de $\sqrt{f(x)}$. Sethi (1963) a proposé une méthode de détermination des bornes par calcul infinitésimal au moyen des équations

$$\frac{(x_h - \mu_h)^2 + \sigma_h^2}{\sigma_h} = \frac{(x_{h+1} - \mu_{h+1})^2 + \sigma_{h+1}^2}{\sigma_{h+1}}$$

pour une distribution continue standard ressemblant à la population étudiée.

La comparaison de certaines méthodes classiques d'approximation révèle que la méthode d'Ekman et celle de Dalenius et Hodges donnent systématiquement de bons résultats (voir Cochran 1961, Hess, Sethi et Balakrishnan 1966, Murthy 1967), mais que la seconde est plus commode et plus facile à appliquer (voir Nicoloni 2001).

Unnithan (1978) a proposé une méthode itérative s'appuyant sur la méthode de Newton modifiée de Shanno pour déterminer les bornes de strate qui donnent lieu à un

1. Mohammad G. Mostafa Khan, Department of Statistics, Iowa State University, Ames, Iowa 50011, É.-U., (M.G.M. Khan était en congé de la School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji) ; Niraj Nand et Nesar Ahmad, School of Computing, Information and Mathematical Sciences, University of the South Pacific, Suva, Fiji.

minimum local de la variance sous la répartition de Neyman, si une solution initiale appropriée est choisie. Il est prouvé que cette méthode est plus rapide que la méthode itérative de Dalenius et Hodges. Plus tard, Unnithan et Nair (1995) ont donné une méthode pour sélectionner un point de départ approprié pour la méthode de Newton modifiée qui permet d'aboutir à un minimum global de la variance.

Lavallée et Hidiroglou (1988) ont proposé un algorithme pour construire les bornes de strate pour un échantillon stratifié et la répartition des unités échantillonnées sans certitude selon une méthode de puissance. Hidiroglou et Srinath (1993) ont présenté une forme plus générale de l'algorithme, qui, lorsqu'on affecte différentes valeurs aux paramètres opérationnels, produit une répartition de puissance, une répartition de Neyman ou une combinaison de ces deux répartitions. Sweet et Sigman (1995), ainsi que Rivest (2002) ont revu l'algorithme de Lavallée et Hidiroglou et proposé leurs versions modifiées dans lesquelles sont intégrées les différentes relations entre la stratification et les variables étudiées. Detlefsen et Veum (1991) ont étudié l'algorithme de Lavallée et Hidiroglou pour plusieurs strates et observé que la convergence de l'algorithme était lente ou inexistante. Ils ont également découvert que différents points de départ aboutissaient à des BOS différentes pour la même population.

Niemiro (1999) a proposé une méthode de recherche aléatoire dans le problème de stratification, mais l'algorithme ne garantissait pas d'aboutir à un optimum global. De surcroît, si la population était grande, il fonctionnait mal, car il nécessitait un trop grand nombre d'étapes d'itération (voir Kozak 2004).

Nicolini (2001) a proposé une méthode, nommée *méthode des classes naturelles* (MCN), en opposition à celle, plus utilisée, de Dalenius et Hodges, mais ni l'une ni l'autre de ces méthodes ne s'est avérée plus efficace que l'autre.

Lednicki et Wieczorkowski (2003) ont présenté une méthode de stratification fondée sur la méthode du simplexe de Nelder et Mead (1965). Plus tard, Kozak (2004) a présenté l'algorithme de recherche aléatoire modifié en tant que méthode de stratification optimale. L'algorithme de Kozak était plus rapide et efficace que celui de Rivest et celui de Lednicki et Wieczorkowski, mais il n'était pas possible de garantir qu'il mène à un optimum global.

Bühler et Deutler (1975) ont formulé le problème de la détermination des BOS comme un problème d'optimisation pouvant être résolu par une méthode de programmation dynamique. Cette approche est également utilisée par Lavallée (1987, 1988) pour trouver les BOS qui diviseraient le domaine de population de deux variables de stratification en sous-ensembles distincts, de telle façon que la précision des variables d'intérêt soit maximisée.

Khan, Khan et Ahsan (2002) ont considéré le problème de la découverte des BOS comme étant équivalant au problème de détermination des largeurs optimales de strate (LOS). Ils ont formulé le problème des LOS comme un problème de programmation mathématique (PPM). En suivant l'approche de programmation dynamique de Bühler et Deutler, ils ont résolu le PPM qui donne la solution exacte si la distribution de fréquence de la variable étudiée est connue et que le nombre de strates est établi d'avance. Khan et coll. (2002) ont appliqué leur méthode pour déterminer les BOS pour une population suivant une loi uniforme et une loi triangulaire droite (*right triangular distribution*). Plus tard, Khan, Najmussehar et Ahsan (2005) ont étendu cette approche de programmation dynamique à la détermination des BOS pour une variable étudiée exponentielle.

Dans le présent article, nous discutons du problème de détermination des BOS pour des variables étudiées suivant une loi triangulaire ou normale standard. Vu que ces problèmes sont équivalents à des problèmes de détermination des LOS, nous les formulons comme des problèmes de programmation mathématique et les résolvons en suivant l'approche de programmation dynamique de Bühler et Deutler. Les PPM formulés minimisent la variance du paramètre de population estimé dans le cas d'une répartition de Neyman sous la contrainte que la somme totale des largeurs des strates est égale à l'étendue totale de la distribution de la variable étudiée. À la section 2, nous passons en revue l'approche de programmation dynamique proposée par Bühler et Deutler (1975). À la section 3, nous donnons les détails de la formulation des problèmes de détermination des LOS comme des problèmes de programmation mathématique. À la section 4, nous discutons de l'algorithme fondé sur la technique de programmation dynamique utilisée pour résoudre les PPM. À la section 5, nous illustrons les calculs de l'algorithme au moyen d'exemples numériques. Enfin, à la section 6, nous comparons les résultats obtenus par la méthode de programmation dynamique et par la méthode de la cumulative de $\sqrt{f}$ de Dalenius et Hodges (1959) en nous basant sur un exemple tiré d'une population suivant une loi normale. Cette comparaison montre que la méthode de programmation dynamique proposée est plus efficace que la méthode de la cumulative de $\sqrt{f}$.

2. Détermination des BOS par des méthodes de programmation dynamique : une revue de l'approche de Bühler et Deutler

Soit X une variable étudiée aléatoire, discrète ou continue, dont la densité de probabilité est $f(x)$, $a \le x \le b$. Pour estimer la moyenne de population μ à l'aide d'un

échantillon stratifié, X est partitionnée en L strates $[a, x_1], (x_1, x_2], ..., (x_{L-1}, b]$ telles que

$$a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b. \quad (1)$$

Supposons que, pour la strate h $(h = 1, 2, ..., L)$, qui contient N_h unités, on sélectionne un échantillon de taille n_h contenant les unités y_{hj} $(h = 1, 2, ..., L; j = 1, 2, ..., n_h)$. Alors, la moyenne stratifiée $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ est une estimation sans biais de μ dont la variance est

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h \sigma_h^2 \left(\frac{W_h}{n_h} - \frac{1}{N} \right), \quad (2)$$

où $W_h = N_h/N$, $\bar{x}_h = 1/n_h \sum_{j=1}^{n_h} y_{hj}$, $\sigma_h^2 = [1/(N_h - 1)] \times \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2$ et $\mu_h = 1/N_h \sum_{j=1}^{N_h} y_{hj}$.

Quand la fonction de fréquence $f(x)$ est connue, les valeurs de W_h et σ_h dans (2) peuvent être obtenues par

$$W_h = \int_{x_{h-1}}^{x_h} f(x)\, dx, \quad (3)$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x)\, dx - \mu_h^2, \quad (4)$$

où

$$\mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx \quad (5)$$

est la moyenne et (x_{h-1}, x_h) sont les bornes de la h^{e} strate.

Alors, (2) se lit comme étant la fonction des bornes de strate et des tailles d'échantillon, autrement dit

$$V(\bar{x}_{st}) = V(\bar{x}_{st} | x_1, ..., x_{L-1}, n_1, ..., n_L).$$

Si les n_h sont fixes, l'objectif de la stratification optimale est de déterminer les bornes de strate $(x_1, ..., x_{L-1})$ telles que $V(\bar{x}_{st})$ est minimale. En outre, si les fractions d'échantillonnage n_h/N_h sont faibles ou que l'échantillonnage est fait avec remise, nous obtenons les problèmes d'optimisation qui suivent, selon le type de répartition de la taille d'échantillon totale $(n = \sum_{h=1}^{L} n_h)$ entre les strates.

1. Répartition proportionnelle $(n_h = n \cdot W_h)$

$$\text{Minimiser} \sum_{h=1}^{L} W_h \sigma_h^2$$

sous la contrainte

$$a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b \quad (6)$$

2. Répartition égale $(n_h = n/L)$

$$\text{Minimiser} \sum_{h=1}^{L} W_h^2 \sigma_h^2$$

sous la contrainte

$$a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b \quad (7)$$

3. Répartition de Neyman $(n_h = n \cdot W_h \sigma_h / \sum_{h=1}^{L} W_h \sigma_h)$

$$\text{Minimiser} \sum_{h=1}^{L} W_h \sigma_h$$

sous la contrainte

$$a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b. \quad (8)$$

Les problèmes (6) à (8) ont la structure suivante :

$$\text{Minimiser} \sum_{h=1}^{L} \phi_h(x_{h-1}, x_h),$$

sous la contrainte

$$a = x_0 \le x_1 \le x_2 \le, ..., \le x_{L-1} \le x_L = b. \quad (9)$$

Bühler et Deutler (1975) ont proposé une méthode d'optimisation récurrente pour résoudre (9) en utilisant une méthode de programmation dynamique comme il suit.

Considérons un problème d'optimisation ayant la structure spéciale :

$$\text{Minimiser} \sum_{h=1}^{m} u_h(z_{h-1}, y_h),$$

sous les contraintes

$$z_h = v_h(z_{h-1}, y_h),$$
$$z_h \in Z_h,$$
$$y_h \in S_h(z_{h-1}),$$
$$z_0 = z'; \; h = 1, 2, ..., m, \quad (10)$$

où m = nombre de degrés, u_h = fonctions de rendement de degré, v_h = fonctions de transformation de degré, Z_h = espaces d'états, S_h = espaces de décisions et z' = état initial. Alors, nous pouvons utiliser une méthode de programmation dynamique en appliquant le principe d'optimalité de Bellman (Bellman 1957) pour résoudre (10).

Si $m = L$, $z_0 = a$, $z_L = b$, $Z_h = [a, b]$, $Z_{h-1} = [a, b - y_h]$, $S_h(z_{h-1}) = [0, b - z_{h-1}]$ avec $z_{h-1} \in Z_{h-1}$, $u_h(z_{h-1}, y_h) = \phi_h(z_{h-1}, y_h + z_{h-1})$ avec $y_h \in S_h(z_{h-1})$, $v_h(z_{h-1}, y_h) = y_h + z_{h-1}$, alors (10) se transforme en le problème suivant :

$$\text{Minimiser} \sum_{h=1}^{L} \phi_h(z_{h-1}, y_h + z_{h-1}),$$

sous les contraintes

$$z_h = y_h + z_{h-1},$$
$$z_h \in [a, b],$$
$$y_h \in [0, b - z_{h-1}],$$
$$z_0 = a, \; z_L = b; \; h = 1, 2, ..., L. \quad (11)$$

Le problème (11) et le problème (9) sont équivalents, car ils comportent les résultats suivants :

1. Si $(x_1^*, ..., x_{L-1}^*)$ est une solution optimale de (9), alors $y_h^* = x_h^* - x_{h-1}^*$, $z_h^* = x_h^*$ est un optimum de (11).
2. Si $y_h^*(h = 1, ..., L)$, $z_h^*(h = 1, ..., L-1)$ est une solution optimale de (11), alors $x_h^* = z_h^*(h = 1, ..., L-1)$ est une solution optimale de (9).

Si $\Phi_h(z_{h-1})$ est la valeur optimale de la fonction objectif au degré h avec l'état disponible z_{h-1}, alors l'équation récurrente rétrograde pour résoudre (11) en utilisant une méthode de programmation dynamique est donnée par

$$\Phi_h(z_{h-1}) = \min[\phi_h(z_{h-1}, y_h + z_{h-1}) + \Phi_{h+1}(z_h) \mid z_h = y_h + z_{h-1}] \qquad (12)$$

sur $y_h \in S_h(z_{h-1})$ avec au départ $\Phi_{L+1} \equiv 0$.

3. Formulation du problème de détermination des LOS comme un PPM

À la présente section, nous étendons l'approche de Bühler et Deutler discutée plus haut au cas d'une variable étudiée dont la densité de probabilité $f(x)$ est continue. Le problème (11) est transformé en un problème équivalent de détermination des LOS en considérant $y_h = z_h - z_{h-1} = x_h - x_{h-1}$ comme les largeurs de strates, puis en construisant la fonction objectif et les contraintes comme des fonctions de y_h. Le PPM est traité comme un problème de décision à plusieurs degrés dans lequel, à chaque degré, la valeur de la LOS, donc de la BOS pour une strate, est déterminée en utilisant une méthode de programmation dynamique avec une équation récurrente progressive.

Soit $f(x)$ la fonction de probabilité, et x_0 et x_L, la plus petite et la plus grande valeurs de x. Si la moyenne de population est estimée sous une répartition de Neyman, le problème de la détermination des bornes des strates consiste à découper l'étendue,

$$x_L - x_0 = d, \qquad (13)$$

aux points intermédiaires $x_1 \le x_2 \le, ..., \le x_{L-1}$, de sorte que dans (8), $\sum_{h=1}^{L} W_h \sigma_h$ soit minimale.

Considérons que $f(x)$ comprend n fonctions linéaires ou non linéaires continues par morceaux telles que :

$$f(x) = \begin{cases} g_1(x); & x_0 = a_0 \le x \le a_1, \\ g_2(x); & a_1 < x \le a_2, \\ \quad \vdots & \\ g_n(x); & a_{n-1} < x \le a_n = x_L. \end{cases} \qquad (14)$$

Supposons aussi que, parmi les L strates, l_i est le nombre de strates qui doivent être formées sous la densité de probabilité $g_i(x)$; $i = 1, 2, ..., n$ et que $\sum_{i=1}^{n} l_i = L$.

Si $f(x)$ dans (14) est intégrable, en utilisant les expressions (3), (4) et (5), nous obtenons W_h, σ_h^2 et μ_h sous forme d'une fonction des bornes x_h et x_{h-1}. Donc, dans (8), la fonction objectif peut être exprimée comme une fonction des bornes x_h et x_{h-1} uniquement. Soit

$$\phi_h(x_h, x_{h-1}) = W_h \sigma_h.$$

Notons que la fonction susmentionnée possède des valeurs différentes pour les diverses densités de probabilité dans (14).

Donc, le problème (8) peut être traité comme un problème d'optimisation en vue de trouver $x_1, x_2, ..., x_{L-1}$ tel qu'il est énoncé dans (9).

Soit $y_h = x_h - x_{h-1} \ge 0$ la largeur de la h^{e} $(h = 1, 2, ..., L)$ strate.

Partant de la définition susmentionnée de y_h, l'étendue de la distribution donnée par (13) est exprimée comme la fonction des largeurs de strate :

$$\sum_{h=1}^{L} y_h = \sum_{h=1}^{L} (x_h - x_{h-1}) = x_L - x_0 = d. \qquad (15)$$

La k^{e} borne de strate x_k ; $k = 1, 2, ..., L-1$ est alors exprimée par :

$$x_k = x_0 + y_1 + y_2 + \ ... \ + y_k$$
$$= x_{k-1} + y_k,$$

qui est une fonction de la k^{e} largeur de strate et de la $(k-1)^{\text{e}}$ borne de strate.

Si nous considérons $z_k = x_k$ et ajoutons (15) comme contrainte, le problème (11) peut être réécrit sous la forme d'un problème équivalent de détermination des LOS :

$$\text{Minimiser } \sum_{h=1}^{L} \phi_h(y_h, x_{h-1}),$$
$$\text{sous les contraintes}$$
$$\sum_{h=1}^{L} y_h = d,$$
$$y_h \ge 0; \ h = 1, 2, ..., L. \qquad (16)$$

Au départ, x_0 est connue. Par conséquent, le premier terme, c'est-à-dire $\phi_1(y_1, x_0)$, de la fonction objectif du PPM (16) est une fonction de y_1 uniquement. Une fois que y_1 est connue, le point de stratification suivant $x_1 = x_0 + y_1$ sera connu et le deuxième terme de la fonction objectif $\phi_2(y_2, x_1)$ deviendra une fonction de y_2 uniquement.

Par conséquent, en énonçant la fonction objectif comme une fonction de y_h uniquement, le PPM (16) s'exprime par :

$$\text{Minimiser} \quad \sum_{h=1}^{L} \phi_h(y_h),$$

sous les contraintes

$$\sum_{h=1}^{L} y_h = d,$$

$$y_h \geq 0\,;\ h = 1,\ 2,\ ...,\ L. \tag{17}$$

Les sections 3.1 et 3.2 illustrent la formulation du problème de la détermination des LOS comme un PPM pour les variables étudiées de loi triangulaire et de loi normale standard, respectivement.

3.1 PPM pour la loi triangulaire

Posons que la variable étudiée x suit la loi triangulaire sur l'intervalle [a, b] avec la densité de probabilité :

$$f(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)}\ ; & a \leq x \leq c \\ \dfrac{2(b-x)}{(b-a)(b-c)}\ ; & c < x \leq b, \end{cases} \tag{18}$$

où a est un paramètre d'emplacement, b est un paramètre d'échelle et c est le paramètre de forme.

Elle possède deux fonctions par morceaux.

Quand $a \leq x \leq c,$ en partant de (18) et en utilisant (3), (5) et (4), nous obtenons W_h, μ_h, et σ_h^2 comme il suit :

$$W_h = \frac{y_h\,(y_h + 2a_h)}{(b-a)(c-a)}, \tag{19}$$

$$\mu_h = \frac{\dfrac{2}{3}y_h^2 + 2y_h x_{h-1} - a y_h + 2a_h x_{h-1}}{y_h + 2a_h},$$

et

$$\sigma_h^2 = \frac{y_h^2\,[\,y_h^2 + 6a_h y_h + 6a_h^2\,]}{18\,(y_h + 2a_h)^2}, \tag{20}$$

où $y_h = x_h - x_{h-1}$, $a_h = x_{h-1} - a$ et $a \leq x_{h-1} \leq x_h \leq c.$

Donc, de (19) et (20), il découle que

$$W_h\,\sigma_h = \frac{y_h^2\sqrt{y_h^2 + 6a_h\,y_h + 6a_h^2}}{3\sqrt{2}(b-a)(c-a)}. \tag{21}$$

De même, quand $c < x \leq b,$ en partant de (18) et en utilisant (3), (5) et (4), nous pouvons démontrer que

$$W_h = \frac{y_h(2b_h - y_h)}{(b-a)(b-c)}, \tag{22}$$

$$\mu_h = \frac{3b_h y_h - 3y_h x_{h-1} + 6b_h x_{h-1} - 2y_h^2}{3(2b_h - y_h)},$$

et

$$\sigma_h^2 = \frac{y_h^2\,(6b_h^2 - 6b_h y_h + y_h^2)}{18\,(2b_h - y_h)^2}, \tag{23}$$

où $y_h = x_h - x_{h-1}$, $b_h = b - x_{h-1}$ et $c < x_{h-1} \leq x_h \leq b.$

Donc, de (22) et (23), il découle que

$$W_h\sigma_h = \frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{3\sqrt{2}(b-a)(b-c)}. \tag{24}$$

Soit λ_1 et λ_2 la dernière et la première strates formées sous la première et la seconde fonctions par morceaux de (18), respectivement. Si toute strate (disons l) tombe sous les deux fonctions, λ_1 et λ_2 ne sont pas considérées comme étant deux strates distinctes, mais comme les fractions de la même l^{e} strate. Alors, en utilisant (21) et (24), le PPM (17) peut être exprimé comme le problème de détermination des LOS pour la variable étudiée dont la fonction de fréquence est triangulaire, sous la forme :

$$\text{Minimiser} \left\{ \sum_{h=1}^{\lambda_1} \frac{y_h^2\sqrt{y_h^2 + 6a_h y_h + 6a_h^2}}{3\sqrt{2}(b-a)(c-a)} + \sum_{h=\lambda_2}^{L} \frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{3\sqrt{2}(b-a)(b-c)} \right\},$$

$$\text{sous les contraintes} \quad \sum_{h=1}^{L} y_h = d,$$

$$\text{et} \quad y_h \geq 0\,;\ h = 1,\ 2,\ ...,\ L, \tag{25}$$

où $d = b - a.$

3.2 PPM pour la loi normale

La variable étudiée x suit une loi normale standard si sa fonction de densité de probabilité est donnée par

$$f(x) = \frac{1}{\sqrt{2\pi}}\ \exp\left(-\frac{x^2}{2}\right); \quad -\infty < x < \infty.$$

Comme à la section 3.1, en utilisant les définition (3), (5) et (4), nous pouvons voir que

$$W_h = \frac{erf\left(\dfrac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\dfrac{x_{h-1}}{\sqrt{2}}\right)}{2}, \tag{26}$$

$$\mu_h = \frac{\sqrt{2}\left[\exp\left(-\dfrac{x_{h-1}^2}{2}\right) - \exp\left(-\dfrac{(y_h + x_{h-1})^2}{2}\right)\right]}{\sqrt{\pi}\left[erf\left(\dfrac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\dfrac{x_{h-1}}{\sqrt{2}}\right)\right]}$$

et

$$\sigma_h^2 = \left\{ \sqrt{2\pi} \left[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) \right.\right.$$

$$- (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right)$$

$$\left. + (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]$$

$$+ \pi \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2$$

$$\left. - 2 \left[\exp\left(-\frac{x_{h-1}^2}{2} \right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) \right]^2 \right\}$$

$$\div \pi \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2, \qquad (27)$$

où $y_h = x_h - x_{h-1}$, $erf(x_h) - erf(x_{h-1}) = (2/\sqrt{\pi}) \times \int_{x_{h-1}}^{x_h} \exp(-u^2)\, du$ et $h = 1, 2, ..., L$.

Par conséquent, en utilisant les valeurs données par (26) et (27), nous pouvons exprimer le PPM (17) sous la forme

$$\text{Minimiser} \sum_{h=1}^{L} \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \left[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) \right.\right.$$

$$- (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right)$$

$$- x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right)$$

$$\left. + (y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]$$

$$+ \frac{1}{4} \left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}} \right) - erf\left(\frac{x_{h-1}}{\sqrt{2}} \right) \right]^2$$

$$\left. - \frac{1}{2\pi} \left[\exp\left(-\frac{x_{h-1}^2}{2} \right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2} \right) \right]^2 \right\}$$

sous les contraintes

$$\sum_{h=1}^{L} y_h = d$$

$$y_h \geq 0 ; \; h = 1, 2, ..., L. \qquad (28)$$

4. L'algorithme fondé sur la méthode de programmation dynamique

Le PPM (17) est un problème de décision à plusieurs degrés dans lequel la fonction objectif et les contraintes peuvent être séparées de y_h, ce qui nous permet d'utiliser une méthode de programmation dynamique telle que celle illustrée par Bühler et Deutler (1975) pour le problème (11).

Considérons le sous-problème suivant de (17) pour les $k (< L)$ premières strates :

$$\text{Minimiser} \sum_{h=1}^{k} \phi_h(y_h),$$

sous les contraintes

$$\sum_{h=1}^{k} y_h = d_k,$$

$$y_h \geq 0 ; \; h = 1, 2, ..., k, \qquad (29)$$

où $d_k < d$ est la largeur totale disponible pour la division en k strates ou la valeur d'état au degré k. Notons que $d_k = d$ pour $k = L$.

Les fonctions de transformation sont données par

$$\begin{aligned} d_k &= y_1 + y_2 + ... + y_k, \\ d_{k-1} &= y_1 + y_2 + ... + y_{k-1} = d_k - y_k, \\ d_{k-2} &= y_1 + y_2 + ... + y_{k-2} = d_{k-1} - y_{k-1}, \\ &\vdots \qquad \vdots \\ d_2 &= y_1 + y_2 = d_3 - y_3, \\ d_1 &= y_1 = d_2 - y_2. \end{aligned}$$

Soit $\Phi_k(d_k)$ la valeur minimale de la fonction objectif de (29), c'est-à-dire

$$\Phi_k(d_k) = \min \left[\sum_{h=1}^{k} \phi_h(y_h) \,\middle|\, \sum_{h=1}^{k} y_h = d_k, \right.$$

$$\left. \text{et} \quad y_h \geq 0; h = 1, 2, ..., k \right].$$

En utilisant la définition susmentionnée de $\Phi_k(d_k)$, le PPM (17) équivaut à calculer $\Phi_L(d)$ par récurrence en trouvant $\Phi_k(d_k)$ pour $k = 1, 2, ..., L$ et $0 \leq d_k \leq d$.

Nous pouvons écrire :

$$\Phi_k(d_k) = \min \left[\phi_k(y_k) + \sum_{h=1}^{k-1} \phi_h(y_h) \,\middle|\, \sum_{h=1}^{k-1} y_h = d_k - y_k, \right.$$

$$\left. \text{et} \;\; y_h \geq 0; \; h = 1, 2, ..., k-1. \right]$$

Pour une valeur fixe de y_k ; $0 \leq y_k \leq d_k$,

$$\Phi_k(d_k) = \phi_k(y_k) + \min\left[\sum_{h=1}^{k-1}\phi_h(y_h) \,\middle|\, \sum_{h=1}^{k-1} y_h = d_k - y_k, \text{ et } y_h \geq 0; h = 1, 2, ..., k-1\right].$$

En appliquant le principe d'optimalité de Bellman, nous écrivons une récurrence progressive au lieu d'une récurrence rétrograde comme l'ont proposé Bühler et Deutler dans (12) pour appliquer la méthode de programmation dynamique, soit :

$$\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k}\left[\phi_k(y_k) + \Phi_{k-1}(d_k - y_k)\right], \; k \geq 2. \quad (30)$$

Pour le premier degré, c'est-à-dire pour $k = 1$:

$$\Phi_1(d_1) = \phi_1(d_1) \Rightarrow \; y_1^* = d_1, \quad (31)$$

où $y_1^* = d_1$ est la largeur optimale de la première strate. Les relations (30) et (31) sont résolues par récurrence pour chaque $k = 1, 2, ..., L$ et $0 \leq d_k \leq d$, et nous obtenons $\Phi_L(d)$. Partant de $\Phi_L(d)$, la largeur optimale de la L^{e} strate, nous obtenons y_L^*. À partir de $\Phi_{L-1}(d - y_L^*)$, la largeur optimale de la $(L-1)^{\text{e}}$ strate, nous obtenons y_{L-1}^*, et ainsi de suite jusqu'à ce que nous obtenions y_1^*.

Soulignons que, selon la fonction par morceaux donnée dans (14) sous laquelle la strate est formée, $\phi_k(y_k)$ dans (30) prendra une valeur différente pour chaque y_k, comme il suit :

$$y_k = x_k - x_{k-1} \leq a_i - a_0, \text{ pour un } i \; (i = 1, 2, ..., n)$$

et

$$x_k \in [a_{i-1}, a_i], \qquad \text{pour un } i \; (i = 1, 2, ..., n).$$

5. Exemples numériques

À la présente section, nous présentons le détail des calculs de l'algorithme exposé à la section 4 pour les PPM (25) et (28).

5.1 Loi triangulaire

Supposons que $a = x_0 = 0$, $c = 1$ et $b = x_L = 2$. Cela implique que $d = x_L - x_0 = 2$ et le PPM (25) s'exprime sous la forme :

$$\text{Minimiser} \left\{\sum_{h=1}^{\lambda_1} \frac{y_h^2\sqrt{y_h^2 + 6a_h y_h + 6a_h^2}}{6\sqrt{2}} + \sum_{h=\lambda_2}^{L} \frac{y_h^2\sqrt{6b_h^2 - 6b_h y_h + y_h^2}}{6\sqrt{2}}\right\},$$

$$\text{sous les contraintes } \sum_{h=1}^{L} y_h = 2,$$

$$\text{et } y_h \geq 0; \; h = 1, 2, ..., L, \quad (32)$$

où $a_h = x_{h-1}$ et $b_h = 2 - x_{h-1}$.

En utilisant (30) et (31), nous pouvons énoncer les équations récurrentes pour la résolution du PPM (32) de la façon suivante :

Pour le premier degré $(k = 1)$

$$\Phi_1(d_1) = \frac{d_1^3}{6\sqrt{2}} \quad \text{à} \quad y_1 = d_1, \quad (33)$$

et pour les autres degrés $(k \geq 2)$

$$\Phi_k(d_k) = \begin{cases} \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6a_k y_k + 6a_k^2}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] \\ \qquad \text{si } \; 0 \leq d_k \leq 1, \\ \min\left[\dfrac{y_k^2\sqrt{6b_k^2 - 6b_k y_k + y_k^2}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] \\ \qquad \text{si } \; 1 < d_k \leq 2, \end{cases} \quad (34)$$

où le minimum est calculé sur $0 \leq y_k \leq d_k$, $a_k = x_{k-1} = d_k - y_k$ et $b_k = 2 - x_{k-1} = 2 - d_k + y_k$.

En remplaçant a_k et b_k par ces valeurs dans (34), nous obtenons

$$\Phi_k(d_k) = \begin{cases} \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6(d_k - y_k)d_k}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] \\ \qquad \text{si } \; 0 \leq d_k \leq 1, \\ \min\left[\dfrac{y_k^2\sqrt{y_k^2 + 6(2 - d_k + y_k)(2 - d_k)}}{6\sqrt{2}} + \Phi_{k-1}(d_k - y_k)\right] \\ \qquad \text{si } \; 1 < d_k \leq 2, \end{cases} \quad (35)$$

où le minimum est calculé sur $0 \leq y_k \leq d_k$.

Puis, en résolvant les équations récurrentes (33) et (35) à l'aide d'un programme informatique développé pour l'algorithme donné à la section 4, nous obtenons les LOS. Le tableau 1 donne les résultats des largeurs optimales de strate y_h^* et, donc, les bornes optimales de strate x_h^*, ainsi que les valeurs de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h)$ pour $L = 2, 3, 4, 5$ et 6.

Tableau 1
Largeurs et bornes optimales de strate pour la loi triangulaire

Nombre de strates L	Largeurs optimales de strate (LOS) (y_h^*)	Bornes optimales de strate (BOS) $(x_h^* = x_{h-1}^* + y_h^*)$	Valeurs optimales de la fonction objectif $\sum_{h=1}^{L} \phi_h(y_h) = \sum_{h=1}^{L} W_h \sigma_h$
2	$y_1^* = 1{,}000000$ $y_2^* = 1{,}000000$	$x_1^* = 1{,}000000$	0,2357022604
3	$y_1^* = 0{,}838081$ $y_2^* = 0{,}411608$ $y_3^* = 0{,}750311$	$x_1^* = 0{,}838081$ $x_2^* = 1{,}249689$	0,1655523797
4	$y_1^* = 0{,}645751$ $y_2^* = 0{,}354249$ $y_3^* = 0{,}354249$ $y_4^* = 0{,}645751$	$x_1^* = 0{,}645751$ $x_2^* = 1{,}000000$ $x_3^* = 1{,}354249$	0,1226262641
5	$y_1^* = 0{,}582819$ $y_2^* = 0{,}319725$ $y_3^* = 0{,}252176$ $y_4^* = 0{,}299439$ $y_5^* = 0{,}545841$	$x_1^* = 0{,}582819$ $x_2^* = 0{,}902544$ $x_3^* = 1{,}154720$ $x_4^* = 1{,}454159$	0,0998893913
6	$y_1^* = 0{,}497369$ $y_2^* = 0{,}272849$ $y_3^* = 0{,}229782$ $y_4^* = 0{,}229782$ $y_5^* = 0{,}272849$ $y_6^* = 0{,}497369$	$x_1^* = 0{,}497369$ $x_2^* = 0{,}770218$ $x_3^* = 1{,}000000$ $x_4^* = 1{,}229782$ $x_5^* = 1{,}502631$	0,0829362498

5.2 Loi normale

Posons que x suit la loi normale standard dans l'intervalle (x_0, x_L). Pour les besoins de l'exemple, nous supposons que $x_0 = -4$ et $x_L = 4$. Alors, $d = 8$, ce qui donne le PPM (28) sous la forme :

$$\text{Minimize} \sum_{h=1}^{L} \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \times \right. \right.$$

$$\left[x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2}\right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right) \right.$$

$$-(y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right)$$

$$-x_{h-1} \exp\left(-\frac{x_{h-1}^2}{2}\right) erf\left(\frac{x_{h-1}}{\sqrt{2}}\right)$$

$$\left. +(y_h + x_{h-1}) \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) erf\left(\frac{x_{h-1}}{\sqrt{2}}\right) \right]$$

$$+\frac{1}{4}\left[erf\left(\frac{y_h + x_{h-1}}{\sqrt{2}}\right) - erf\left(\frac{x_{h-1}}{\sqrt{2}}\right) \right]^2$$

$$\left. \left. -\frac{1}{2\pi}\left[\exp\left(-\frac{x_{h-1}^2}{2}\right) - \exp\left(-\frac{(y_h + x_{h-1})^2}{2}\right) \right]^2 \right\} \right\},$$

sous les contraintes $\sum_{h=1}^{L} y_h = 8,$

et $y_h \geq 0\,;\ h = 1,\ 2,\ ...,\ L.$ (36)

Nous avons

$$\begin{aligned} x_{k-1} &= x_0 + y_1 + y_2 + ... + y_{k-1} \\ &= -4 + y_1 + y_2 + ... + y_{k-1} \\ &= d_{k-1} - 4 \\ &= d_k - y_k - 4. \end{aligned}$$

En introduisant cette valeur de x_{k-1} par substitution dans (36) et en utilisant (30) et (31), nous obtenons les équations récurrentes pour la résolution du PPM (36) comme il suit.

Pour le premier degré $(k = 1)$:

$$\Phi_1(d_1) = \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \left[-\exp\left(-\frac{1}{2}\right) erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right) \right.\right.\right.$$

$$-(d_1 - 4) \exp\left(-\frac{(d_1 - 4)^2}{2}\right) erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right)$$

$$+ \exp\left(-\frac{1}{2}\right) erf\left(-\frac{1}{\sqrt{2}}\right)$$

$$\left. + (d_1 - 4) \exp\left(-\frac{(d_1 - 4)^2}{2}\right) erf\left(-\frac{1}{\sqrt{2}}\right) \right]$$

$$+\frac{1}{4}\left[erf\left(\frac{(d_1 - 4)}{\sqrt{2}}\right) - erf\left(-\frac{1}{\sqrt{2}}\right) \right]^2$$

$$\left. \left. -\frac{1}{2\pi}\left[\exp\left(-\frac{1}{2}\right) - \exp\left(-\frac{(d_1 - 4)^2}{2}\right) \right]^2 \right\} \right\} \quad (37)$$

à $y_1 = d_1,$

et pour les degrés $k \geq 2$:

$$\Phi_k(d_k) = \min_{0 \leq y_k \leq d_k} \left\{ \text{Sqrt} \left\{ \frac{1}{2\sqrt{2\pi}} \times \right. \right.$$

$$\left[(d_k - y_k - 4) \exp\left(-\frac{(d_k - y_k - 4)^2}{2}\right) erf\left(\frac{d_k - 4}{\sqrt{2}}\right) \right.$$

$$-(d_k - 4) \exp\left(-\frac{(d_k - 4)^2}{2}\right) erf\left(\frac{d_k - 4}{\sqrt{2}}\right)$$

$$\left. -(d_k - y_k - 4) \exp\left(-\frac{(d_k - y_k - 4)^2}{2}\right) erf\left(\frac{(d_k - y_k - 4)}{\sqrt{2}}\right) \right.$$

$$+ (d_k - 4)\exp\left(-\frac{(d_k-4)^2}{2}\right) erf\left(\frac{(d_k - y_k - 4)}{\sqrt{2}}\right)\Bigg]$$

$$+ \frac{1}{4}\left[erf\left(\frac{d_k-4}{\sqrt{2}}\right) - erf\left(\frac{(d_k-y_k-4)}{\sqrt{2}}\right)\right]^2$$

$$- \frac{1}{2\pi}\left[\exp\left(-\frac{(d_k-y_k-4)^2}{2}\right) - \exp\left(-\frac{(d_k-4)^2}{2}\right)\right]^2\Bigg\}$$

$$+ \Phi_{k-1}(d_k - y_k)\Bigg\}. \qquad (38)$$

En résolvant les équations récurrentes (37) et (38), nous obtenons les largeurs optimales de strate y_h^* et, donc, les bornes optimales de strate x_h^*. Le tableau 2 donne ces résultats, ainsi que les valeurs de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h)$ pour $L = 2, 3, 4, 5$ et 6.

6. Discussion

À la présente section, nous effectuons une étude numérique de l'efficacité de la méthode de programmation dynamique comparativement à la méthode de la cumulative de $\sqrt{f}$ de Dalenius et Hodges, qui est celle utilisée le plus fréquemment et la mieux connue. Pour cela, nous avons produit des données de taille $N = 10\,000$ pour une population de densité de probabilité normale standard $f(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$, que nous avons groupées en 19 classes égales. Au tableau 3, nous donnons les fréquences de classe dans la colonne 2 et leurs racines cumulatives dans la colonne 3.

Dans cet exemple, la plus petite et la plus grande valeurs de x sont $x_0 = -3{,}98$ et $x_L = 3{,}62$, respectivement. Par conséquent, l'étendue de la distribution est $d = 7{,}60$.

Pour déterminer les BOS pour cette distribution, nous utilisons la méthode de la cumulative de $\sqrt{f}$ ainsi que la méthode de programmation dynamique. Pour chaque $L = 2, 3, 4, 5$ et 6, nous calculons la variance $\sum_{h=1}^{L} W_h\sigma_h$, que nous utilisons pour évaluer l'efficacité des deux méthodes de stratification. Les résultats de cette comparaison sont donnés au tableau 4. L'examen de la dernière colonne du tableau montre que les BOS obtenues par la méthode de programmation dynamique sont plus efficaces pour tout $L = 1, 2, ..., 6$. En outre, l'efficacité de la méthode de la cumulative de $\sqrt{f}$ dépend du choix initial du nombre de classes, mais il n'existe aucun fondement théorique donnant le meilleur nombre de classes (voir Hedlin 2000).

Tableau 2
Largeurs et bornes optimales de strate pour la loi normale standard

Nombre de strates L	Largeurs optimales de strate (LOS) (y_h^*)	Bornes optimales de strate (BOS) $(x_h^* = x_{h-1}^* + y_h^*)$	Valeurs optimales de la fonction objectif $\sum_{h=1}^{L}\phi_h(y_h) = \sum_{h=1}^{L} W_h\sigma_h$
2	$y_1^* = 4{,}000000$ $y_2^* = 4{,}000000$	$x_1^* = 0{,}000000$	0,6021710931
3	$y_1^* = 3{,}450300$ $y_2^* = 1{,}099400$ $y_3^* = 3{,}450300$	$x_1^* = -0{,}549700$ $x_2^* = 0{,}549700$	0,4265717619
4	$y_1^* = 3{,}124570$ $y_2^* = 0{,}875430$ $y_3^* = 0{,}875430$ $y_4^* = 3{,}124570$	$x_1^* = -0{,}875430$ $x_2^* = 0{,}000000$ $x_3^* = 0{,}875430$	0,3297899642
5	$y_1^* = 2{,}896360$ $y_2^* = 0{,}767900$ $y_3^* = 0{,}671480$ $y_4^* = 0{,}767900$ $y_5^* = 2{,}896360$	$x_1^* = -1{,}103640$ $x_2^* = -0{,}335740$ $x_3^* = 0{,}335740$ $x_4^* = 1{,}103640$	0,2686646379
6	$y_1^* = 2{,}722440$ $y_2^* = 0{,}702200$ $y_3^* = 0{,}575360$ $y_4^* = 0{,}575360$ $y_5^* = 0{,}702200$ $y_6^* = 2{,}722440$	$x_1^* = -1{,}277560$ $x_2^* = -0{,}575360$ $x_3^* = 0{,}000000$ $x_4^* = 0{,}575360$ $x_5^* = 1{,}277560$ $x_6^* = 4{,}000000$	0,2265979522

Tableau 3
Distribution des fréquences de x et cum $\sqrt{f(x)}$

Classe	Fréquence $f(x)$	Cum $\sqrt{f(x)}$
(-3,98)-(-3,58)	2	1,4
(-3,58)-(-3,18)	6	3,8
(-3,18)-(-2,78)	23	8,6
(-2,78)-(-2,38)	59	16,3
(-2,38)-(-1,98)	155	28,7
(-1,98)-(-1,58)	296	45,9
(-1,58)-(-1,18)	630	71,0
(-1,18)-(-0,783)	1 015	102,9
(-0,783)-(-0,383)	1 361	139,8
(-0,383)-0,017	1 551	179,2
0,017-0,417	1 495	217,9
0,417-0,817	1 315	254,2
0,817-1,22	1 003	285,9
1,22-1,62	613	310,7
1,62-2,02	285	327,6
2,02-2,42	128	338,9
2,42-2,82	38	345,1
2,82-3,22	18	349,3
3,22-3,62	7	351,9

Enfin, les autres méthodes décrites dans la littérature, telles que celles de Aoyama (1954), Ekman (1959) et Sethi (1963), sont avant tout des méthodes classiques appliquées en vue d'obtenir des bornes de strate approximatives. De nombreux auteurs, dont Unithan (1978), Lavallée et Hidiroglou (1988), Sweet et Sigman (1995), et Rivest (2002), ont proposé des méthodes itératives. Ces dernières requièrent des solutions initiales approximatives. En outre, il n'existe aucune garantie qu'une méthode itérative donnera

le minimum global en l'absence d'une solution initiale approximative appropriée et la fonction de variance possède plus d'un minimum local. L'avantage de la méthode de programmation dynamique tient au fait qu'elle donne le minimum global de la fonction objectif et qu'elle ne nécessite aucune solution initiale approximative.

Tableau 4
Efficacité relative de la méthode de programmation dynamique

L	Méthode cum $\sqrt{f}$		Méthode de programmation dynamique		Efficacité relative
	BOS	$\sum_{h=1}^{L} W_h \sigma_h$	BOS	$\sum_{h=1}^{L} W_h \sigma_h$	
2	-0,017	0,60131	-0,00034	0,60126	100,00832
3	-0,783	0,43177	-0,55015	0,42576	101,41159
	0,417		0,54884		
4	-0,783	0,33067	-0,87593	0,32905	100,49233
	-0,017		-0,00081		
	0,817		0,87395		
5	-1,18	0,27066	-1,10418	0,26799	100,99631
	-3,83		-0,33656		
	0,417		0,33452		
	1,22		1,10147		
6	-1,18	0,24242	-1,27813	0,22598	107,27498
	-0,783		-0,57619		
	-0,017		-0,00115		
	0,417		0,57369		
	1,22		1,27462		

Remerciements

Les auteurs remercient le rédacteur associé et les examinateurs de leurs suggestions et commentaires précieux en vue d'améliorer le manuscrit

Bibliographie

Aoyama, H. (1954). A study of stratified random sampling. *Annals of the Institute of Statistical Mathematics*, 6, 1-36.

Bellman, R.E. (1957). *Dynamic Programming*. Princetown University Press, New Jersey.

Bühler, W., et Deutler, T. (1975). Optimal stratification and grouping by dynamic programming. *Metrika*, 22, 161-175.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 38, 2, 345-358.

Dalenius, T. (1950). The problem of optimum stratification-II. *Skand. Aktuartidskr*, 33, 203-213.

Dalenius, T., et Gurney, M. (1951). The problem of optimum stratification. *Skand. Aktuartidskr*, 34, 133-148.

Dalenius, T., et Hodges, J.L. (1959). Minimum variance stratification. *Journal of the Americal Statistical Association*, 54, 88-101.

Detlefsen, R.E., et Veum, C.S. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 214-219.

Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Annals of the Institute of Statistical Mathematics*, 30, 1131-1134.

Hansen, M.H., et Hurwitz, W.N. (1953). On the theory of sampling from finite population. *The Annals of Mathematical Statistics*, 14, 333-362.

Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16, 15-29.

Hess, I., Sethi, V.K. et Balakrishnan, T.R. (1966). Stratification: A practical investigation. *Journal of the Americal Statistical Association*, 61, 71-90.

Hidiroglou, M.A., et Srinath, K.P. (1993). Problems associated with designing subannual bussiness surveys. *Journal of Bussiness and Economic Statistics*, 11, 397-405.

Khan, E.A., Khan, M.G.M. et Ahsan, M.J. (2002). Optimum stratification: A mathematical programming approach. *Culcutta Statistical Association Bulletin*, 52 (spécial), 205-208.

Khan, M.G.M., Najmussehar, et Ahsan, M.J. (2005). Optimum stratification for exponential study variable under Neyman allocation. *Journal of Indian Society of Agricultural Statistics*, 59(2), 146-150.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 5, 797-806.

Lavallée, P. (1987). Some contributions to optimal stratification. Thèse de maîtrise, Carleton University, Ottawa, Canada.

Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Virginia, 646-651.

Lavallée, P., et Hidiroglou, M. (1988). Sur la stratification de populations asymétriques. *Techniques d'enquête*, 14, 35-45.

Lednicki, B., et Wieczorkowski, R. (2003). Optimal stratification and sample allocation between subpopulations and strata. *Statistics in Transition*, 6, 287-306.

Mahalanobis, P.C. (1952). Some aspects of the design of sample surveys. *Sankhyā*, 12, 1-7.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Nelder, J.A., et Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7, 308-313.

Nicolini, G. (2001). A method to define strata boundaries. *Departmental Working Papers 2001-01*, Department of Economics, University of Milan, Italie (www.economia.unimi.it/pubb/wp83.pdf).

Niemiro, W. (1999). Konstrukcja optymalnej stratyfikacja metoda poszukiwan losowych. (Optimal Stratification using Random Search Method). *Wiadomosci Statystyczne*, 10, 1-9.

Rivest, L.-P. (2002). Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises. *Techniques d'enquête*, 28, 207-214.

Sethi, V.K. (1963). A note on optimum stratification of population for estimating the population mean. *Australian Journal of Statistics*, 5, 20-33.

Sweet, E.M., et Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 491-496.

Unnithan, V.K.G. (1978). The minimum variance boundry points of stratification. *Sankhyā*, 40, C, 60-72.

Unnithan, V.K.G., et Nair, N.U. (1995). Minimum variance stratification. *Communications in Statistics*, 24(1), 275-284.

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 237-245
Statistique Canada, N° 12-001-X au catalogue

Optimisation multi-objective pour une répartition optimale dans l'échantillonnage stratifié multivarié

José A. Díaz-García et Liliana Ulloa Cortez [1]

Résumé

Le présent article porte sur le traitement de la répartition optimale dans l'échantillonnage stratifié multivarié comme une optimisation matricielle non linéaire en nombres entiers. Un problème non linéaire d'optimisation multi-objectifs en nombres entiers est étudié à titre de cas particulier. Enfin, un exemple détaillé, y compris certaines méthodes proposées, est donné à la fin de l'exposé.

Mots clés : Échantillonnage stratifié multivarié ; répartition optimale ; optimisation multi-objectifs.

1. Introduction

L'échantillonnage probabiliste est l'un des volets de la statistique utilisés le plus fréquemment dans tous les domaines de recherche scientifique. Une bonne méthode d'échantillonnage d'une population donne lieu à une extraction appropriée de données utiles qui permettent de bien comprendre les aspects importants de la population. L'échantillonnage stratifié est l'une des méthodes classiques d'obtention de ce genre d'information. Il s'appuie sur le calcul de la taille des échantillons de strate, qui peut se faire selon diverses méthodes. Cependant, on a montré que la répartition optimale était une approche utile. La répartition optimale est considérée comme un problème d'optimisation non linéaire dans lequel la fonction objectif est la variance sous une contrainte de coût, ou inversement. Habituellement, on le résout en utilisant l'inégalité de Cauchy-Schwarz (Stuart 1954) citée dans Cochran (1977), ou la méthode du multiplicateur de Lagrange (voir à ce sujet Sukhatme, Sukhatme, Sukhatme et Asok 1984).

La théorie classique de l'échantillonnage s'appuie sur une variable ou un paramètre de décision unique ; par exemple, dans notre cas, l'échantillonnage stratifié univarié repose sur l'étude d'un paramètre, la taille d'échantillon et la répartition de cet échantillon entre les strates (voir Cochran 1977, Sukhatme et coll. 1984, et Thompson 1997). En outre, dans le contexte de l'échantillonnage stratifié, certains auteurs ont proposé des approches multivariées tenant compte de diverses caractéristiques pour déterminer la taille de l'échantillon et sa répartition entre les strates (voir, entre autres, Sukhatme et coll. 1984 et Arthanari et Dodge 1981).

Si l'on procède à une répartition optimale et que la fonction de coût est la fonction objectif, sous certaines contraintes de variance dans les différentes caractéristiques, le problème peut se réduire à une question de programmation mathématique classique et, à ce propos, il existe deux approches bien connues, à savoir celle d'Arthanari et Dodge (1981), d'un point de vue déterministe et celle de Prékopa (1978), d'un point de vue stochastique. Dans le second cas, le problème peut être résolu en utilisant n'importe quelle méthode présentée dans Díaz-García et Garay (2007).

Autrement, si nous souhaitons minimiser les variances sous la contrainte d'une fonction de coût ou d'une taille donnée d'échantillon, nous pouvons adopter plusieurs approches pour résoudre le problème (voir Sukhatme et coll. 1984). Toutefois, les approches susmentionnées ne résolvent pas le problème du suréchantillonnage, autrement dit les situations où la taille d'échantillon dans une ou plusieurs strates est plus grande que la taille de la strate ; de surcroît, les tailles d'échantillon obtenues ne sont pas des nombres entiers et doivent être approximées. En outre, comme nous le verrons, toutes les approches publiées antérieurement dans ce domaine sont des cas particuliers de la méthode d'optimisation multi-objectifs. Si ces difficultés pouvaient être surmontées, nous aurions un aperçu formel et une théorie unifiée pour résoudre le problème de la répartition optimale sous échantillonnage stratifié multivarié et nous pourrions prendre en considération toute la littérature (théorique ainsi que pratique) sur l'optimisation multi-objectifs et les questions connexes.

Dans le présent article, nous étudions la répartition optimale dans l'échantillonnage stratifié multivarié comme un problème non linéaire d'optimisation matricielle en nombres entiers sous la contrainte d'une fonction de coût ou d'une taille d'échantillon donnée. En formulant certaines hypothèses, nous proposons un moyen de résoudre le problème, par plusieurs méthodes particulières (voir la sous-section 3.1). Le deuxième objectif de l'article est lié au fait suivant : si nous définissons une fonction vectorielle particulière de la fonction objectif du problème d'optimisation matricielle, alors à la sous-section 3.2, nous montrons que la

1. José A. Díaz-García et Liliana Ulloa Cortez, Department of Statistics and Computation, Universidad Autónoma Agraria Antonio Narro, 25350 Buenavista, Saltillo, Coahuila, Méxique. Courriel : jadiaz@uaaan.mx and lullcor@uaan.mx.

répartition optimale dans l'échantillonnage stratifié multivarié peut aussi être étudiée comme un problème non linéaire d'optimisation multi-objectifs en nombres entiers. Aux sous-sections 3.2.1 et 3.2.2, nous proposons diverses méthodes pour résoudre ces problèmes. Enfin, à la section 4, nous appliquons certaines des méthodes décrites à un exemple numérique concernant la sylviculture.

2. Échantillonnage stratifié multivarié

Considérons une population de taille N, divisée en H sous-populations (strates). Nous souhaitons trouver un échantillon représentatif de taille n et une répartition optimale entre les strates satisfaisant les exigences suivantes : i) minimiser la variance de la moyenne estimée sous une contrainte budgétaire ou ii) minimiser le coût sous une contrainte de variance ; il s'agit du problème classique de répartition optimale dans l'échantillonnage stratifié univarié (voir Cochran 1977, Sukhatme et coll. 1984, et Thompson 1997). Par contre, si nous considérons plus d'une caractéristique (variable), le problème est celui d'une répartition optimale dans l'échantillonnage stratifié multivarié. Afin d'obtenir une expression formelle du problème de répartition optimale dans l'échantillonnage stratifié, considérons la notation qui suit.

2.1 Notation

L'indice inférieur $h = 1, 2, \ldots, H$ désigne la strate, $i = 1, 2, \ldots, N_h$, l'unité dans la strate h et $j = 1, 2, \ldots, G$, la caractéristique (variable). En outre :

N_h — Nombre total d'unités dans la strate h

n_h — Nombre d'unités provenant de l'échantillon dans la strate h

y_{hi}^j — Valeur obtenue pour la i^{e} unité dans la strate h de la j^{e} caractéristique

$\mathbf{n} = (n_1, \ldots, n_H)'$ — Vecteur du nombre d'unités dans l'échantillon

$W_h = \dfrac{N_h}{N}$ — Taille relative de la strate h

$\overline{Y}_h^j = \dfrac{\sum_{i=1}^{N_h} y_{hi}^j}{N_h}$ — Moyenne de population dans la strate h de la j^{e} caractéristique

$\overline{y}_h^j = \dfrac{\sum_{i=1}^{n_h} y_{hi}^j}{n_h}$ — Moyenne d'échantillon dans la strate h de la j^{e} caractéristique

$\overline{\mathbf{y}}_h = (\overline{y}_h^1, \ldots, \overline{y}_h^G)'$ — Vecteur des moyennes d'échantillon dans la strate h

$\overline{y}_{\text{ST}}^j = \sum_{h=1}^{H} W_h \overline{y}_h^j$ — Estimateur de la moyenne de population dans l'échantillonnage stratifié multivarié pour la j^{e} caractéristique

$\overline{\mathbf{y}}_{\text{ST}} = (\overline{y}_{\text{ST}}^1, \ldots, \overline{y}_{\text{ST}}^G)'$ — Estimateur du vecteur des moyennes de population dans l'échantillonnage stratifié multivarié

$s_{h_{jk}}$ — Covariance d'échantillon dans la strate h des j^{e} et k^{e} caractéristiques, où

$$s_{h_{jk}} = \frac{\sum_{i=1}^{n_h} (y_{hi}^j - \overline{y}_h^j)(y_{hi}^k - \overline{y}_h^k)}{n_h - 1}, \text{ et}$$

$$s_{h_{jj}} = s_{hj}^2 \frac{\sum_{i=1}^{n_h} (y_{hi}^j - \overline{y}_h^j)^2}{n_h - 1}$$

$\text{Cov}(\overline{\mathbf{y}}_{\text{ST}})$ — Matrice de variance-covariance de $\overline{\mathbf{y}}_{\text{ST}}$

$\widehat{\text{Cov}}(\overline{\mathbf{y}}_{\text{ST}})$ — Estimateur de la matrice de variance-covariance de $\overline{\mathbf{y}}_{\text{ST}}$, où $\widehat{\text{Cov}}(\overline{\mathbf{y}}_{\text{ST}}) \equiv \widehat{\text{Cov}(\overline{\mathbf{y}}_{\text{ST}})}$ et

$$\widehat{\text{Cov}}(\overline{\mathbf{y}}_{\text{ST}}) = \begin{pmatrix} \widehat{\text{Var}}(\overline{y}_{\text{ST}}^1) & \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^1, \overline{y}_{\text{ST}}^2) & \cdots & \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^1, \overline{y}_{\text{ST}}^G) \\ \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^2, \overline{y}_{\text{ST}}^1) & \widehat{\text{Var}}(\overline{y}_{\text{ST}}^2) & \cdots & \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^2, \overline{y}_{\text{ST}}^G) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^G, \overline{y}_{\text{ST}}^1) & \widehat{\text{Cov}}(\overline{y}_{\text{ST}}^G, \overline{y}_{\text{ST}}^2) & \cdots & \widehat{\text{Var}}(\overline{y}_{\text{ST}}^G) \end{pmatrix}$$

$\widehat{\text{Cov}}(\overline{y}_{\text{ST}}^j, \overline{y}_{\text{ST}}^k)$ — Covariance estimée de $\overline{y}_{\text{ST}}^j$ et $\overline{y}_{\text{ST}}^k$, où

$$\widehat{\text{Cov}}(\overline{y}_{\text{ST}}^i, \overline{y}_{\text{ST}}^j) \equiv \widehat{\text{Cov}(\overline{y}_{\text{ST}}^j, \overline{y}_{\text{ST}}^k)}, \quad \text{avec}$$

$$\widehat{\text{Cov}}(\overline{y}_{\text{ST}}^j, \overline{y}_{\text{ST}}^k) = \sum_{h=1}^{H} \frac{W_h^2 s_{h_{jk}}}{n_h} - \sum_{h=1}^{H} \frac{W_h s_{h_{jk}}}{N}, \quad \text{et}$$

$$\widehat{\text{Cov}}(\overline{y}_{\text{ST}}^j, \overline{y}_{\text{ST}}^j) \equiv \widehat{\text{Var}}(\overline{y}_{\text{ST}}^j) = \sum_{h=1}^{H} \frac{W_h^2 s_{hj}^2}{n_h} - \sum_{h=1}^{H} \frac{W_h s_{hj}^2}{N}$$

c_h — Coût par unité d'échantillonnage dans la strate h

Enfin, soit $\mathbf{V}_u(\bar{\mathbf{y}}_{ST}) \in \Re^G$, de sorte que $\mathbf{V}_u(\bar{\mathbf{y}}_{ST}) = (\widehat{\text{Var}}(\bar{y}_{ST}^1), \ldots, \widehat{\text{Var}}(\bar{y}_{ST}^G))'$, où, si $\mathbf{a} \in \Re^G$, $\mathbf{a}'$ désigne la transposée de $\mathbf{a}$.

3. Nouvelle approche du problème de répartition optimale dans l'échantillonnage stratifié multivarié

À la présente section, nous proposons d'envisager la répartition optimale dans l'échantillonnage stratifié multivarié comme un problème d'optimisation matricielle pour lequel nous étudions un certain nombre de solutions possibles. Nous constatons que le problème d'optimisation multi-objectifs est un cas particulier d'une optimisation matricielle. Dans le même ordre d'idée, nous notons que la répartition optimale dans l'échantillonnage stratifié multivarié peut être considérée comme un problème d'optimisation multi-objectifs. Dans chaque cas, les solutions respectives se calculent facilement.

3.1 Optimisation matricielle

Formellement, la répartition optimale dans l'échantillonnage stratifié peut être étudiée en résolvant le problème d'optimisation matricielle non linéaire suivant :

$$\begin{aligned} &\min_{\mathbf{n}} \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}) \\ &\text{sous la contrainte} \\ &\mathbf{c}'\mathbf{n} + c_0 = C, \end{aligned} \tag{1}$$

où C est le coût total, c_0 est un coût fixe et $\mathbf{c}' = (c_1, \ldots, c_H)$.

Notons que les solutions proposées pour le problème (1) prennent des valeurs réelles; or, les tailles d'échantillon n_h doivent être des entiers. Nous devons également résoudre le problème du suréchantillonnage, c'est-à-dire les situations où $n_h \geq N_h$ pour au moins certaines strates h (voir Arvanitis et Afonja 1971). Afin de surmonter ces deux complications, nous proposons, au lieu de (1), l'approche alternative suivante :

$$\begin{aligned} &\min_{\mathbf{n}} \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}) \\ &\text{sous les contraintes} \\ &\mathbf{c}'\mathbf{n} + c_0 = C \\ &2 \leq n_h \leq N_h, h = 1, 2, \ldots, H \\ &n_n \in \mathbb{N}, \end{aligned} \tag{2}$$

où $\mathbb{N}$ désigne l'ensemble de nombres naturels.

Manifestement, la difficulté que pose l'expression du problème de cette façon tient à la définition de la signification du minimum d'une fonction matricielle. L'idée de minimiser une fonction matricielle, en particulier la matrice de variance-covariance, a été étudiée dans divers domaines de la théorie statistique. Ainsi, pour déterminer les estimateurs par la régression d'un modèle linéaire général multivarié, on minimise le déterminant ou la trace des sommes des carrés et des sommes des matrices produits des erreurs (voir Giri 1977). De même, le choix ou la comparaison de certains modèles de plans expérimentaux se fait par minimisation d'une fonction de la matrice de variance-covariance des estimateurs des traitements (voir Khuri et Cornell 1987 et Azaïs et Druilhet 1997).

Heureusement, il est possible de réduire le problème de minimisation matricielle non linéaire (2) à un problème de minimisation univariée non linéaire en tenant compte des éléments suivants (notons que la méthode décrite ici n'est que l'une des diverses options possibles (voir Ríos, Ríos Insua et Ríos Insua 1989 et Miettinen 1999). Observons que $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST})$ est une fonction explicite de $\mathbf{n}$ et qu'elle doit donc être représentée par $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}) \equiv \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$. En outre, supposons que $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}))$ est une matrice définie positive pour tout $\mathbf{n}$, $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n})) > \mathbf{0}$. Soit, maintenant, $\mathbf{n}_1$ et $\mathbf{n}_2$ deux valeurs possibles du vecteur $\mathbf{n}$ et soit $\mathbf{B} = \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) - \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2))$. Nous disons que

$$\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) < \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2)) \Leftrightarrow \mathbf{B} < \mathbf{0}, \tag{3}$$

c'est-à-dire si la matrice $\mathbf{B}$ est définie négative. En outre, notons que $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1))$ et $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2))$ sont diagonalisables. Alors, soit $D_{\mathbf{n}_1}$ et $D_{\mathbf{n}_2}$ les matrices diagonales associées à $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1))$ et $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2))$, respectivement, avec $D_{\mathbf{n}_1} = \text{diag}(\alpha_1, \ldots, \alpha_G), \alpha_1 > \cdots > \alpha_G > 0$ et $D_{\mathbf{n}_2} = \text{diag}(\tau_1, \ldots, \tau_G), \tau_1 > \cdots > \tau_G > 0$, où α_j et τ_j désignent les valeurs propres de $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1))$ et $\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2))$, respectivement. Donc, l'expression (3) peut aussi être présentée sous la forme :

$$\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) < \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2)) \Leftrightarrow D_{\mathbf{n}_1} - D_{\mathbf{n}_2} < \mathbf{0},$$

c'est-à-dire

$$\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) < \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2)) \Leftrightarrow \alpha_j - \tau_j < 0 \quad \underset{j=1,\ldots,G}{}$$

et

$$\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) \neq \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2)), \tag{4}$$

qui définit un ordre de Pareto faible (voir Steuer 1986, Ríos et coll. 1989 et Miettinen 1999). Alors, d'après Steuer (1986), Ríos et coll. (1989) et Miettinen (1999), il existe une fonction $\mathbf{f}: S \to \Re$, telle que

$$\begin{aligned} &\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1)) < \widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2)) \\ &\Leftrightarrow \mathbf{f}\left(\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_1))\right) < \mathbf{f}\left(\widehat{\text{Cov}}(\bar{\mathbf{y}}_{ST}(\mathbf{n}_2))\right). \end{aligned} \tag{5}$$

où $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})) \in S \subset \Re^{G(G+1)/2}$ et S est l'ensemble de matrices définies positives. Partant de (5), Steuer (1986), Ríos et coll. (1989) et Miettinen (1999) prouvent que le problème de minimisation matricielle non linéaire (2) se réduit au problème de minimisation univariée non linéaire suivant :

$$\min_{\mathbf{n}} \mathbf{f}\left(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}})\right)$$

sous les contraintes

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \tag{6}$$

$$2 \le n_h \le N_h,\ h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N};$$

Malheureusement ou heureusement, la fonction $\mathbf{f}(\cdot)$ n'est pas unique. Par exemple, dans d'autres contextes statistiques, nous voyons les fonctions $\mathbf{f}(\cdot)$ suivantes utilisées fréquemment (voir Giri 1977) :

1. la trace de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$; $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \mathrm{tr}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})))$;
2. le déterminant de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$; $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \left|\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))\right|$;
3. la somme de tous les éléments de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$;

$$\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \sum_{j,k=1}^{G} \widehat{\mathrm{Cov}}(\bar{y}_{\mathrm{ST}}^{j}, \bar{y}_{\mathrm{ST}}^{k})\ ;$$

4. $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \lambda_{\max}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})))$, où $\lambda_{\max}$ est la valeur propre maximale de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$;
5. $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \lambda_{\min}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})))$, où $\lambda_{\min}$ est la valeur propre minimale de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$;
6. $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \lambda_j(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n})))$, où λ_j est la j^{e} valeur propre de la matrice $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$, entre autres.

En particulier, Dalenius (1957) a étudié le problème (6) quand $\mathbf{f}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))) = \left|\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))\right|$, autrement dit la minimisation de la variance généralisée $\left|\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))\right|$; voir aussi Arvanitis et Afonja (1971).

3.2 Optimisation multi-objectifs

Considérons maintenant la fonction vectorielle $\mathbf{F}: S \to \Re^G$, telle que $\mathbf{F}(\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}})) = \mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}})$. Un autre moyen d'établir le problème (2) est

$$\min_{\mathbf{n}} \mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}) = \min_{\mathbf{n}} \begin{pmatrix} \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{1}) \\ \vdots \\ \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{G}) \end{pmatrix}$$

sous les contraintes

$$\mathbf{c}'\mathbf{n} + c_0 = C \tag{7}$$

$$2 \le n_h \le N_h,\ h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N},$$

qui est un problème non linéaire de l'optimisation multi-objectifs en nombres entiers (voir Steuer 1986, Ríos et coll. 1989 et Miettinen 1999).

Dans le contexte de l'échantillonnage, soulignons que, dans les problèmes d'optimisation multi-objectifs, il existe rarement un point $\mathbf{n}^*$ considéré comme un optimum, autrement dit, peu de cas satisfont à l'exigence que $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{j}(\mathbf{n}^*))$ soit minimale pour tout $j = 1, \ldots, G$. Cela justifie la notion qui suit du point de Pareto (qui est défini plus faiblement qu'un point optimal) :

> *Nous disons que* $\mathbf{V}_u^*(\bar{\mathbf{y}}_{\mathrm{ST}})$ *est un point de Pareto de* $\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}})$ *s'il n'existe aucun autre point* $\mathbf{V}_u^1(\bar{\mathbf{y}}_{\mathrm{ST}})$ tel que $\mathbf{V}_u^1(\bar{\mathbf{y}}_{\mathrm{ST}}) \le \mathbf{V}_u^*(\bar{\mathbf{y}}_{\mathrm{ST}})$, *c'est-à-dire pour tout* j, $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{j_1}) \le \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{j_*})$ *et* $\mathbf{V}_u^1(\bar{\mathbf{y}}_{\mathrm{ST}}) \ne \mathbf{V}_u^*(\bar{\mathbf{y}}_{\mathrm{ST}})$.

Les critères d'existence de points de Pareto d'un problème d'optimisation multi-objectifs sont établis dans Ríos et coll. (1989) et dans Miettinen (1999). En particulier, nous avons :

> *Soit* $\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}): \Re^H \to \Re^G$ *et considérons un compact non vide* $\mathcal{N} \subset \mathbb{N}^H$ *tel que* $\mathcal{N}$ *est l'ensemble de toutes les valeurs possibles de* $\mathbf{n}$ *déterminées par les contraintes en* (*7*). *Si* $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{j})$ *est une fonction semi-continue supérieurement pour chaque* $j = 1, \ldots, G$, *alors le problème* (*7*) *possède une solution de Pareto optimale.*

Par ailleurs, Steuer (1986), Ríos et coll. (1989) et Miettinen (1999) ont étudié l'extension de l'optimisation scalaire (*conditions de Kuhn-Tucker*) au cas vectoriel. En particulier, ils ont proposé les conditions nécessaires pour les solutions de Pareto, qui deviennent des conditions suffisantes si N est convexe, si les fonctions $\widehat{\mathrm{Var}}(\bar{\mathbf{y}}_{\mathrm{ST}}^{j})$, $j = 1, \ldots, G$ sont convexes et si les multiplicateurs généralisés de Lagrange δ_j, associés à chaque fonction $\widehat{\mathrm{Var}}(\bar{\mathbf{y}}_{\mathrm{ST}}^{j})$, sont positifs, $\delta_j > 0$ pour tout j. Notons que les résultats susmentionnés pour l'existence d'une solution de Pareto et pour les conditions de Kuhn-Tucker sont valides quand les variables

n_h sont continues, pour tout $h = 1, \ldots, H$. Cependant, il faut se souvenir qu'afin d'obtenir la solution d'un problème d'optimisation en nombres entiers, il est normal d'émettre l'hypothèse initiale que ce genre de problème est un problème d'optimisation continue. Pour commencer, nous calculons la solution du problème d'optimisation continue, puis, par des méthodes heuristiques ou de séparation et évaluation (*branch-and-bound*), nous passons à la résolution du problème d'optimisation en nombres entiers. Dans ce contexte, pour l'optimisation en nombres entiers, en pratique, il est suffisant de voir que le problème correspondant d'optimisation continue possède une solution de Pareto optimale et de confirmer que l'ensemble $\mathcal{N}$ de toutes les valeurs possibles de $\mathbf{n}$ contient au moins un $\mathbf{n} \in \mathcal{N}$ pour lequel toutes les coordonnées sont des nombres entiers.

Les méthodes de résolution d'un problème d'optimisation multi-objectifs sont basées sur l'information que l'on possède au sujet d'un problème particulier. Trois scénarios sont possibles, selon que l'information est complète, partielle ou nulle (voir Ríos et coll. 1989, Miettinen 1999 et Steuer 1986). Dans un contexte d'échantillonnage stratifié, posséder une information complète signifie que le chercheur connaît suffisamment bien la population pour proposer une fonction de valeur (il s'agit d'une fonction $\phi: \Re^H \to \Re$ telle que, en notant $\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}) \equiv \mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}))$, nous avons que $\min \mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}^*)) < \min \mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}_1)) \Leftrightarrow \phi(\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}^*))) < \phi(\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}(\mathbf{n}_1))), \mathbf{n}^* \neq \mathbf{n}_1$) reflétant l'importance de chaque variance des caractéristiques étudiées, possibilité qui, de nos jours, se rencontre très rarement. En cas d'information partielle, le chercheur connaît très bien la caractéristique principale étudiée et il s'agit d'un appui suffisant pour la recherche. Enfin, en cas d'information nulle, situation qui est la plus fréquente, le chercheur ne possède des renseignements qu'au sujet des estimateurs du paramètre de l'expérience, ce qui lui permet de trouver une solution appropriée.

Faute d'espace, il est impossible de donner une explication exhaustive de toutes les méthodes proposées pour résoudre les problèmes d'optimisation multi-objectifs (7) ; pour une description détaillée, consulter Ríos et coll. (1989), Miettinen (1999) et Steuer (1986). En outre, au lieu des méthodes classiques, certaines méthodes heuristiques permettent d'aborder le problème d'une autre façon (voir Jones et coll. 2002). En guise d'exemple, nous donnons ci-après un aperçu de deux méthodes utilisées fréquemment, la première ayant trait au cas où l'information est complète (fonction de valeur, également appelée fonction d'utilité) et la deuxième, au cas où l'information est nulle (méthode fondée sur les distances).

3.2.1 Fonction de valeur

Comme nous l'avons mentionné plus haut, cette méthode s'applique au cas où l'information est complète si bien que le chercheur est capable de résumer l'importance de toutes les caractéristiques étudiées dans une fonction réelle, comme il est illustré au paragraphe suivant (voir aussi Ríos et coll. 1989, Miettinen 1999 et Steuer 1986, entre autres).

Sous la méthode de la fonction de valeur, le problème (7) s'exprime de la façon suivante :

$$\begin{aligned} &\min_{\mathbf{n}} \phi(\mathbf{V}_u(\bar{y}_{\mathrm{ST}})), \\ &\text{sous les contraintes} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (8) \\ &2 \le n_h \le N_h, h = 1, 2, \ldots, H \\ &n_h \in \mathbb{N}, \end{aligned}$$

où $\phi(\cdot)$ est une fonction scalaire qui résume l'importance de chacune des variances des G caractéristiques.

Il est clair que de nombreuses approches décrites dans la littérature traitant de la répartition optimale dans l'échantillonnage stratifié multivarié, dont la répartition intermédiaire (*compromise allocation*), la répartition intermédiaire minimisant la perte relative totale et la répartition intermédiaire en prenant la moyenne des valeurs optimales (voir Sukhatme et coll. 1984), sont des cas particuliers de la méthode susmentionnée.

Notons que la fonction de valeur $\phi(\cdot)$ peut prendre un nombre infini de formes, ce qui constitue un obstacle fondamental à sa définition. Cependant, certaines fonctions simples ont donné d'excellents résultats dans les applications et peuvent être considérées comme des approches prometteuses. L'une de ces formes particulières est celle de la méthode de pondération. Sous cette approche, le problème (8) peut s'exprimer :

$$\begin{aligned} &\min_{\mathbf{n}} \sum_{j=1}^{G} w_j \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j), \\ &\text{sous les contraintes} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \\ &2 \le n_h \le N_h, h = 1, 2, \ldots, H \\ &n_h \in \mathbb{N}, \end{aligned}$$

de manière que $\sum_{j=1}^{G} w_j = 1, w_j \ge 0 \; \forall \; j = 1, 2, \ldots, G$ où w_j pondère l'importance de chaque caractéristique.

Des diverses méthodes multi-objectifs, nous constatons que la fonction de valeur est, en général, celle qui est appliquée le plus fréquemment, parce que ses propriétés sont celles qui ont été étudiées le plus en détail (voir Ríos et coll. 1989, Miettinen 1999, Steuer 1986, ainsi que les références qu'ils mentionnent).

3.2.2 Méthode fondée sur les distances

Parfois, le chercheur ne possède pas suffisamment d'information préalable au sujet des variables ou a de la difficulté à décider quelles sont les caractéristiques les plus importantes de l'expérience. Le cas échéant, la méthode décrite ici est très utile, parce qu'elle ne nécessite pas beaucoup d'antécédents ; en outre, elle ne requiert qu'un vecteur d'objectifs *idéaux*, qui est déterminé avec l'information nulle exprimée dans le problème (voir Ríos et coll. 1989 et Steuer 1986).

Alors, le problème (7) est résolu en obtenant les valeurs optimales par minimisation de la distance entre la valeur optimale et le vecteur de cibles.

Soit θ_j le point ou objectif idéal pour la fonction objectif $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j)$, $j = 1, \ldots, G$, c'est-à-dire que le vecteur de cibles $\boldsymbol{\Theta}$ est donné par

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_G \end{pmatrix}.$$

Notons que nous pouvons calculer le vecteur de cibles $\boldsymbol{\Theta}$ sans information supplémentaire, ce qui est un des grands avantages de cette méthode. En fait, il se calcule en minimisant individuellement la fonction objectif $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j)$, $j = 1, \ldots, G$, de sorte que le vecteur $\boldsymbol{\Theta}$ est défini comme étant le vecteur des minima individuels de cette fonction, ce que nous obtenons en résolvant les G problèmes de minimisation non linéaires en nombres entiers (voir Rao 1979) suivants :

$$\min_{\mathbf{n}} \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j),$$

sous les contraintes

$$\sum_{h=1}^{H} c_h n_h + c_0 = C$$

$$2 \le n_h \le N_n$$

$$h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N},$$

pour $j = 1, \ldots, G$.

Lorsque le vecteur $\boldsymbol{\Theta}$ est calculé, nous étudions le problème d'optimisation avec la nouvelle fonction objectif, à savoir

$$\min_{\mathbf{n}} d(\mathbf{V}_u(\bar{\mathbf{y}}_{\mathrm{ST}}), \boldsymbol{\Theta})$$

sous les contraintes

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (9)$$

$$2 \le n_h \le N_h, \; h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N},$$

où $d(\cdot, \cdot)$ représente une fonction de distance générale. En particulier, quand nous appliquons le programme (9) à la distance euclidienne, nous obtenons

$$\min_{\mathbf{n}} \sum_{j=1}^{G} \left[\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j) - \theta_j \right]^2$$

sous les contraintes

$$\sum_{h=1}^{H} c_h n_h + c_0 = C$$

$$2 \le n_h \le N_h, \; h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N}.$$

Une autre option est la fonction de distance qui a été proposée par Khuri et Cornell (1987) :

$$\min_{\mathbf{n}} \sum_{j=1}^{G} \frac{\left(\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^j) - \theta_j \right)^2}{\theta_j^2}$$

sous les contraintes

$$\sum_{h=1}^{H} c_h n_h + c_0 = C$$

$$2 \le n_h \le N_h, \; h = 1, 2, \ldots, H$$

$$n_h \in \mathbb{N}.$$

Remarques

1. Soulignons que nous avons utilisé la contrainte de coût $\sum_{h=1}^{H} c_h n_h + c_0 = C$ dans chaque méthode d'optimisation. Cependant, dans certaines situations, nous ne restreignons pas les coûts, mais imposons des contraintes sur la disponibilité des heures-personnes pour exécuter une enquête ou sur le temps total disponible pour exécuter l'enquête, *etc*. Ces limites peuvent être décrites en utilisant l'expression suivante (voir Arthanari et Dodge 1981) :

$$\sum_{h=1}^{H} n_h = n.$$

2. Notons que les méthodes d'optimisation multi-objectifs proposées ici sont générales et qu'elles doivent être ajustées dans certains problèmes ; par exemple, nous ne tenons pas compte de l'unité (grandeur) de chaque variance dans les sommes respectives pour la fonction de valeur. Nous proposons une solution, à savoir remplacer la variance de chaque caractéristique par le coefficient de variation correspondant

$$\sqrt{\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{j})}\,/\,\bar{y}_{\mathrm{ST}}^{j},\ j=1\ldots,G\,;$$

ensuite, il est préférable d'utiliser la distance de Khuri et Cornell plutôt que la distance euclidienne.

3. Il est souhaitable d'envisager d'autres estimateurs qu'un estimateur de la moyenne, par exemple l'estimateur de la moyenne nationale, ou la comparaison de moyennes régionales, *etc*. En particulier, le rédacteur associé a recommandé des estimateurs du type :

$$\hat{\mathbf{T}}=\sum_{h=1}^{H} w_h\bar{\mathbf{y}}_h\in\Re^G \qquad (10)$$

où plusieurs poids w_h pourraient même être utilisés pour la même variable. Par exemple, si l'un des poids w_h est égal à 1, qu'un autre est égal à -1 et que les autres sont égaux à 0, nous pouvons calculer la différence entre les moyennes de deux strates différentes. En général, nous pouvons optimiser le problème (2) en remplaçant la fonction objectif $\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}})$ par toute fonction d'intérêt. Par exemple, nous pourrions utiliser la matrice de variance-covariance estimée $\widehat{\mathrm{Cov}}(\hat{\mathbf{T}})$ de l'estimateur (10), parmi de nombreuses autres options.

4. Exemple numérique

Les données d'entrée sont tirées d'Arvanitis et Afonja (1971) qui décrivent une enquête sur les forêts réalisée dans le comté de Humbolt, en Californie. La population a été subdivisée en neuf strates en fonction du volume de bois d'œuvre par unité de surface déterminé d'après des photographies aériennes. Les deux variables incluses dans l'exemple sont la surface terrière (ST) (en terminologie sylvicole, « surface terrière » s'entend de la surface d'une section perpendiculaire à l'axe longitudinal d'un arbre mesurée à 4,5 pieds du sol) en pieds carrés et le volume net en pieds cubes (Vol.), exprimées toutes deux sur la base d'un acre. Le tableau 1 donne les variances, les covariances et le nombre d'unités dans la strate h.

Pour cet exemple, le problème d'optimisation matricielle sous l'approche (2) est

$$\min_{\mathbf{n}}\begin{pmatrix}\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{1}) & \widehat{\mathrm{Cov}}(\bar{y}_{\mathrm{ST}}^{1},\bar{y}_{\mathrm{ST}}^{2})\\ \widehat{\mathrm{Cov}}(\bar{y}_{\mathrm{ST}}^{2},\bar{y}_{\mathrm{ST}}^{1}) & \widehat{\mathrm{Var}}(\bar{y}_{\mathrm{ST}}^{2})\end{pmatrix}$$

sous les contraintes

$$\sum_{h=1}^{9} n_h=1\,000 \qquad (11)$$

$$2\le n_h\le N_h,\ h=1,\ldots,9$$

$$n_h\in\mathbb{N}.$$

Le tableau 2 donne les solutions de l'optimisation obtenues par certaines des méthodes décrites aux sections 2 et 3. En particulier, nous présentons les solutions pour les approches de la trace, du déterminant, de la fonction de valeur, de la distance euclidienne et de la distance de Khuri et Cornell. Nous incluons aussi la répartition optimale pour chaque caractéristique, pour ST et pour Vol. (les deux premières lignes du tableau 2). Les deux dernières colonnes donnent les valeurs minimales des variances individuelles pour les répartitions optimales produites par chaque méthode, respectivement. Pour calculer les résultats, nous nous sommes servis du logiciel commercial Hyper LINGO/PC, version 6.0 (voir Winston 1995). Les méthodes d'optimisation utilisées par défaut par LINGO pour résoudre les problèmes non linéaires d'optimisation en nombres entiers sont celles du gradient réduit généralisé (GRG) et de la séparation et évaluation (voir Bazaraa et coll. 2006). Suivent certains aspects techniques des calculs : le nombre maximal d'itérations des méthodes présentées au tableau 2 était 1 193 (problème du déterminant) et la durée moyenne d'exécution était d'une seconde pour tous les programmes. Enfin, notons que l'écart le plus important entre les tailles des strates produit par diverses méthodes a été obtenu en minimisant la variance généralisée $|\widehat{\mathrm{Cov}}(\bar{\mathbf{y}}_{\mathrm{ST}})|$. Cela tient, sans aucun doute, au fait qu'il s'agit de la seule des méthodes présentées au tableau 2 qui tient compte de la covariance entre les deux caractéristiques étudiées.

Tableau 1
Variances, covariances et nombre d'unités dans chaque strate

		Variance		
Strate	N_h	ST	Vol.	Covariance
1	11 131	1 557	554 830	28 980
2	65 857	3 575	1 430 600	61 591
3	106 936	3 163	1 997 100	72 369
4	72 872	6 095	5 587 900	166 120
5	78 260	10 470	10 603 000	293 960
6	51 401	8 406	15 828 000	357 300
7	24 050	20 115	26 643 000	663 300
8	46 113	9 718	13 603 000	346 810
9	102 985	2 478	1 061 800	39 872

Tableau 2
Tailles d'échantillon et estimateur des variances pour les diverses répartitions calculées

Répartition	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	$\widehat{\mathbf{Var}}(\bar{y}_{\mathbf{ST}}^{1})$	$\widehat{\mathbf{Var}}(\bar{y}_{\mathbf{ST}}^{2})$
ST	10	94	144	136	191	113	81	109	122	5,591	5 441,105
Vol.	7	62	119	136	200	161	98	134	83	5,953	5 139,531
$\mathrm{tr}\left(\widehat{\mathrm{Cov}}(\bar{y}_{\mathrm{ST}})\right)$	7	62	119	135	200	161	98	134	84	5,591	5 139,531
$\lvert\,\widehat{\mathrm{Cov}}(\bar{y}_{\mathrm{ST}})\rvert$	9	93	128	129	193	123	86	106	133	5,616	5 403,876
Fonction de valeur[a]	7	62	119	135	200	161	98	134	84	5,591	5 139,531
d_E [b]	7	62	119	136	200	160	98	134	84	5,944	5 139,557
d_{KC} [c]	10	86	137	135	192	126	86	115	113	5,613	5 308,11

[a] $w_1 = w_2 = 0,50$

[b] Distance euclidienne

[c] Distance de Khuri et Cornell

5. Conclusion

Il est difficile de proposer des règles en vue de choisir une méthode pour l'optimisation matricielle (2) quand il existe d'importantes différences numériques entre deux d'entre elles. Par exemple, le tableau 2 révèle des résultats contradictoires en ce qui concerne les répartitions optimales et les variances minimales pour les méthodes de la trace et du déterminant. Une situation semblable se présente en ce qui concerne le choix du critère pour les tests d'hypothèse dans le problème MANOVA (voir Giri 1977). En fait, l'existence de critères généraux fondés sur des tests de puissance ne suffit pas pour prendre une décision objective et le choix final dépend des compétences du chercheur.

Cependant, quand le problème de la répartition optimale dans l'échantillonnage stratifié multivarié est considéré comme un problème non linéaire d'optimisation multi-objectifs en nombres entiers, nous pouvons formuler certaines recommandations générales afin de réduire le nombre de méthodes appropriées dans chaque situation. Premièrement, nous devons reconnaître le contexte de recherche du problème (c'est-à-dire information complète, information partielle ou information nulle). Puis, nous pouvons choisir la méthode en fonction de l'information disponible. Il importe de souligner qu'on ne doit appliquer qu'une seule méthode pour arriver à la solution d'un problème de répartition. Par conséquent, les résultats obtenus pour un exemple particulier ne sont comparables que dans le contexte dans lequel cet exemple a été établi.

Remerciements

Les travaux de recherche décrits ici ont été financés en partie par CONACYT-Mexico, subvention de recherche n° 45974-F, 81512 et par IDI-Spain, subvention n° MTM2005-09209. Les auteurs tiennent aussi à remercier les examinateurs et le rédacteur de leurs suggestions très utiles.

Bibliographie

Arthanari, T.S., et Dodge, Y. (1981). *Mathematical Programming in Statistics*. A Wiley-Interscience Publication John Wiley & Sons, Inc.

Arvanitis, L.G., et Afonja, B. (1971). Use of the generalized variance and the gradient projection method in multivariate stratified sampling. *Biometrics*, 27, 119-127.

Azaïs, J.-M., et Druilhet, P. (1997). Optimality of neighbour balanced designs when neighbour effects are neglected. *Journal of Statistical Planning and Inference*, 64, 353-367.

Bazaraa, M.S., Sherali, H.D. et Shetty, C.M. (2006). *Nonlinear Programming*: *Theory and Algorithms*, 3ième édition. Wiley-Interscience, 2006.

Cochran, W.G. (1977). *Técnicas de muestreo*. C.E.C.S.A. México (en espagnol).

Dalenius, T. (1957). *Sampling in Sweden. Contribution to the methods and theories of sample survey practice*. Almqvist and Wicksell, Stockholm.

Díaz-García, J.A., et Garay Tapia, M.M. (2007). Optimum allocation in stratified surveys: Stochastic programming. *Computational Statistics and Data Analysis*, 51, 3016-3026.

Giri, N.C. (1977). *Multivariate statistical inference*. New York : Academic Press.

Jones, D.F., Mirrazavi, S.K. et Tamiz, M. (2002). Multi-objective metaheuristics: An overview of the current state-of-the art. *European Journal of Operational Research*, 137, 1-9.

Khuri, A.I., et Cornell, J.A. (1987) *Response Surface*: *Designs Analysis*. New York : Marcel Dekker, Inc.

Miettinen, K.M. (1999). *Non linear multiobjective optimization*. Kluwer Academic Publishers, Boston.

Prékopa, A. (1978). The use of stochastic programming for the solution of some problems in statistics and probability. Technical Summary report 1983. University of Wisconsin-Madison.

Rao, S.S. (1979). *Optimization theory and applications*. Wiley Eastern Limited.

Ríos, S., Ríos Insua, S., et Ríos Insua, M.J. (1989). *Procesos de decisión Multicriterio*. EUDEMA, Madrid (en espagnol).

Steuer, R.E. (1986). *Multiple criteria optimization*: *Theory*, *computation and applications*. New York : John Wiley & Sons, Inc.

Stuart, A. (1954). A simple presentation of optimum sampling results. *Journal of the Royal Statistical Society*, B, 16, 239-241.

Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. et Asok, C. (1984). *Sampling theory of surveys with application.* State University Press.

Thompson, M.E. (1997). *Theory of sample surveys*. Chapman & Hall.

Winston, W.L. (1995). *Introduction to mathematical programming*: *Applications and algorithms*. Duxbury Press.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 247-259
Statistique Canada, N° 12-001-X au catalogue

Une approche d'échantillonnage équilibré pour des plans de sondage à stratification multidimensionnelle pour l'estimation pour petits domaines

Piero Demetrio Falorsi et Paolo Righi [1]

Résumé

Le présent article décrit une stratégie d'échantillonnage utile pour obtenir une taille d'échantillon planifiée pour des domaines appartenant à différentes partitions de la population et pour garantir que les erreurs d'échantillonnage des estimations de domaine soient inférieures à un seuil donné. La stratégie d'échantillonnage, qui englobe le cas multidomaine multivarié, est avantageuse quand la taille globale d'échantillon est bornée et que, par conséquent, la solution standard consistant à utiliser un échantillon stratifié dont les strates sont obtenues par le recoupement des variables qui définissent les diverses partitions n'est pas faisable, puisque le nombre de strates est plus grand que la taille globale d'échantillon. La stratégie d'échantillonnage proposée est fondée sur l'utilisation d'une méthode d'échantillonnage équilibré et sur une estimation de type GREG. Le principal avantage de la solution est la faisabilité des calculs, laquelle permet de mettre en œuvre facilement une stratégie globale d'estimation pour petits domaines qui tient compte simultanément du plan d'échantillonnage et de l'estimateur, et qui améliore l'efficacité des estimateurs directs de domaine. Les propriétés empiriques de la stratégie d'échantillonnage étudiée sont illustrées au moyen d'une simulation portant sur des données de population réelles et divers estimateurs de domaine.

Mots clés : Planification de la taille d'échantillon pour les petits domaines ; sélection contrôlée ; échantillonnage équilibré.

1. Introduction

Habituellement, le problème des petits domaines est traité par la voie de l'estimation. Toutefois, s'il existe des variables indicatrices de domaine pour chaque unité de la population, d'autres possibilités peuvent être exploitées à l'étape de l'élaboration du plan de sondage. Cette condition est habituellement remplie dans le contexte des enquêtes-entreprises, où les variables indicatrices de domaine sont disponibles dans le registre des entreprises. Comme l'ont souligné Singh, Gambino et Mantel (1994), il est nécessaire d'élaborer une *stratégie globale* de traitement des problèmes liés aux petits domaines, qui englobe la planification du plan d'échantillonnage et les divers aspects de l'estimation. Dans ce cadre, il est essentiel de contrôler la taille d'échantillon pour chaque domaine d'intérêt, afin que chacun soit traité, à l'étape de l'élaboration du plan, comme un domaine planifié pour lequel il est possible de produire des estimations directes avec un niveau de précision préétabli. En général, dans le cas d'une approche d'inférence fondée sur le plan de sondage, la présence d'unités d'échantillonnage dans chaque domaine permet de calculer des estimations de domaine, quoique celles-ci ne soient pas forcément fiables. En outre, dans l'approche fondée sur un modèle ou assistée par modèle, la présence d'unités d'échantillonnage dans chaque domaine d'estimation permet d'utiliser des modèles comportant des effets de petit domaine spécifiques, ce qui donne des estimations plus précises des paramètres d'intérêt au niveau du petit domaine (Lehtonen, Särndal et Veijanen 2003). Marker (1999, 2001) propose, pour traiter les problèmes d'élaboration du plan de sondage dans le contexte des petits domaines, des stratégies d'échantillonnage fondées sur la stratification et le suréchantillonnage, qui accroissent le nombre de petits domaines pour lesquels il est possible de produire des estimations directes exactes. Ces stratégies sont applicables si les domaines sont emboîtés, mais pourraient ne pas l'être quand le but de l'enquête est de produire des estimations pour deux partitions ou plus de la population. Une solution standard en vue d'obtenir des tailles d'échantillon planifiées pour les domaines de deux ou plusieurs partitions consiste à utiliser un échantillon stratifié dans lequel les strates sont déterminées par le recoupement des variables définissant les diverses partitions. Dans la suite de l'exposé, ce plan sera appelé *plan par recoupement*. Cependant, dans de nombreuses situations pratiques, ce type de plan ne convient pas, parce qu'il nécessite la sélection d'au moins un certain nombre d'unités d'échantillonnage aussi grandes que le produit des nombres de catégories des variables de stratification. Cochran (1977, page 124) illustre bien ce problème à l'aide d'un exemple clair dans lequel le plan par recoupement n'est pas réalisable.

Les renseignements généraux qui précèdent sont typiques du contexte des enquêtes-entreprises. Le *Règlement du Conseil* de l'Union européenne concernant les statistiques structurelles sur les entreprises établit que les paramètres

1. Piero Demetrio Falorsi, Institut national de statistique de l'Italie. Courriel : falorsi@istat.it ; Paolo Righi, Institut national de statistique de l'Italie. Courriel : parighi@istat.it.

d'intérêt se rapportent à des domaines d'estimation définis par trois sous-ensembles distincts de partitions de la population d'entreprises. Par exemple, comme le montre l'examen du tableau 1.1, en Italie, le nombre total de domaines d'estimation est égal à 1 821, tandis que le nombre de strates non vides du plan par recoupement est supérieur à 37 000.

Afin de surmonter certains problèmes que posent les plans par recoupement, une stratégie simple consiste à laisser tomber une ou plusieurs variables de stratification ou à regrouper certaines catégories. Cependant, certains domaines planifiés deviennent non planifiés et certains d'entre eux peuvent avoir une taille d'échantillon faible ou nulle.

De nombreuses méthodes ont été proposées dans la littérature afin de contrôler la taille d'échantillon dans toutes les catégories des variables de stratification sans recourir à un plan par recoupement. Ces méthodes, généralement appelées *méthodes de stratification multidimensionnelle (ou multiple*), ont été élaborées selon deux grandes approches, soit i) des schémas fondés sur des carrés latins (Jessen 1970) et ii) des problèmes d'arrondissement contrôlé résolus par programmation linéaire (Lu et Sitter 2002). Les deux approches ont certains inconvénients qui ont limité l'utilisation des méthodes de stratification multidimensionnelle en tant que solution standard pour la planification des plans d'échantillonnage dans le contexte des sondages réels. En effet, comme l'ont décrit Falorsi, Orsini et Righi (2006), dans nombre de ces contextes, il est impossible, par exemple, de mettre en œuvre des schémas fondés sur des carrés latins quand il n'existe aucune unité de population dans une ou plusieurs strates obtenues par recoupement. La principale faiblesse de la programmation linéaire est la complexité des calculs. La stratégie d'échantillonnage examinée dans le présent article ne souffre pas des inconvénients des méthodes susmentionnées et permet de contrôler les tailles d'échantillon pour les domaines d'intérêt, qui sont définis par différentes partitions de la population de référence. En outre, elle garantit que les erreurs d'échantillonnage des estimations de domaine seront inférieures aux seuils fixés.

La stratégie d'échantillonnage proposée est fondée sur l'utilisation de la méthode d'*échantillonnage équilibré* (Deville et Tillé 2004), ainsi que sur l'estimation de type GREG (Lehtonen et coll. 2003). Comme il est montré dans l'étude portant sur des données empiriques illustrée ici et dans Falorsi et Righi (2008), les principaux avantages de cette solution sont la faisabilité des calculs et l'efficacité, c'est-à-dire l'obtention, pour le cas multidomaine multivarié, d'erreurs d'échantillonnage raisonnablement proches de celles définies par les solutions univariées optimales. Cela permet d'appliquer équitablement une stratégie d'estimation pour petits domaines qui tient compte simultanément du plan d'échantillonnage et de l'estimateur, et qui accroît l'efficacité des estimateurs directs de domaine.

Dans certains contextes de sondage, la stratégie d'échantillonnage proposée pourrait définir une taille globale d'échantillon trop grande pour qu'il soit certain que les bornes préétablies des erreurs d'échantillonnage des estimations de domaine directes soient respectées, ce qui pourrait être dû à un trop grand nombres de domaines pour une partition donnée de population. Si la taille globale d'échantillon est bornée par des contraintes budgétaires, la stratégie d'échantillonnage proposée n'est pas toujours réalisable avec des estimateurs directs. Par conséquent, il pourrait être nécessaire d'adopter un estimateur de petit domaine indirect afin de contrôler les erreurs quadratiques moyennes des estimations des domaines des partitions. Cependant, l'approche proposée pourrait être étendue facilement à une stratégie prévoyant l'utilisation de l'estimateur direct et d'estimateurs indirects de petit domaine pour les partitions qui requièrent une taille globale d'échantillon trop grande pour le bornage des erreurs d'échantillonnage.

La présentation de l'article est la suivante. À la section 2, nous énonçons le problème, nous présentons la notation essentielle et nous décrivons la stratégie globale d'échantillonnage. À la section 3, nous donnons les algorithmes appliqués pour trouver les probabilités d'inclusion et les tailles d'échantillon de domaine planifiées correspondantes. Aux sections 4 et 5, nous illustrons des extensions de la stratégie d'échantillonnage. À la section 4, nous étudions le cas où le critère de variance est représenté par la variance anticipée et, à la section 5, nous présentons une extension au cas d'un estimateur indirect de petit domaine simple. À la section 6, nous donnons les principaux résultats d'une étude empirique portant sur une population réelle d'entreprises italiennes. Enfin, à la section 7, nous énonçons brièvement certaines conclusions.

Tableau 1.1
Nombre de domaines de l'enquête italienne pour les statistiques structurelles d'entreprises par partition

Partition	Nombre de domaines
Classe d'activités économiques (classification à 4 chiffres de la NACE Rév. 1)	465
Groupe d'activités économiques (classification à 3 chiffres de la NACE Rév. 1) par classe de taille[1]	395
Division d'activités économiques (classification à 2 chiffres de la NACE Rév. 1) par région[1]	961
Nombre total de domaines d'estimation	1.821

[1] Les classes de taille sont définies en fonction du nombre de personnes employées.
[2] Les régions sont au nombre de 21, y compris les provinces autonomes.

2. La stratégie d'échantillonnage

2.1 Paramètres d'intérêt

Afin de définir formellement le problème, représentons par U une population de N éléments et par b une partition particulière de $U (b = 1, ..., B)$ dans laquelle la b^{e} partition définit M_b domaines non chevauchants différents, U_{bd} $(d = 1, ..., M_b)$, de taille N_{bd} telle que $\sum_{d=1}^{M_b} N_{bd} = N$ et, enfin, représentons par $\sum_{b=1}^{B} M_b = Q$ le nombre global de domaines.

Soit $y_{r,k}$ et ${}_{bd}\delta_k$, respectivement, la valeur de la r^{e} $(r = 1, ..., R)$ variable d'intérêt dans la k^{e} unité de population et l'indicateur d'appartenance au domaine, égal à ${}_{bd}\delta_k = 1$ si $k \in U_{bd}$ et ${}_{bd}\delta_k = 0$, autrement. Supposons que les valeurs de ${}_{bd}\delta_k$ sont connues pour chaque unité de la population.

Les paramètres d'intérêt sont les $M = Q \times R$ totaux de domaine

$$ {}_{bd}t_r = \sum_{k \in U} y_{r,k}\, {}_{bd}\delta_k = \sum_{k \in U_{bd}} y_{r,k} $$

$$ (r = 1, ..., R\,;\, b = 1, ..., B\,;\, d = 1, ..., M_b). \quad (2.1.1) $$

L'expression (2.1.1) définit un problème *multidomaine multivarié*, puisqu'il existe R variables d'intérêt (aspect multivarié) et $Q > 1$ domaines (aspect multidomaine).

2.2 Description concise de la stratégie d'échantillonnage

Supposons qu'afin d'estimer les paramètres ${}_{bd}t_r$, on sélectionne un échantillon s de taille fixe n à partir de la population U avec les probabilités d'inclusion $\pi_k (k \in U)$. Soit $s_{bd} = s \cap U_{bd}$ l'échantillon de n_{bd} unités appartenant au domaine U_{bd} (avec $\sum_{d=1}^{M_b} n_{bd} = n$) où

$$ n_{bd} = \sum_{k \in U_{bd}} \lambda_k = \sum_{k \in U_{bd}} \pi_k, \quad (2.2.1) $$

avec $\lambda_k = 1$ si $k \in s$ et $\lambda_k = 0$ autrement.

L'échantillon est sélectionné par une *méthode de stratification multidimensionnelle* établie dans le cadre d'*échantillonnage équilibré* qui garantit que l'échantillon sélectionné respecte les *équations d'équilibrage* suivantes

$$ \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}} \quad (2.2.2) $$

où $\hat{t}_{\mathbf{z},ht} = \sum_{k \in U} \mathbf{z}_k \lambda_k a_k$ désigne les estimations d'Horvitz-Thompson de $t_{\mathbf{z}} = \sum_{k \in U} \mathbf{z}_k$, avec $\mathbf{z}_k$ représentant un vecteur de valeurs des variables auxiliaires connues pour chaque unité de population à l'étape de l'élaboration du plan de sondage et $a_k = 1/\pi_k$. La spécification appropriée des vecteurs $\mathbf{z}_k$ permet de s'assurer que les tailles d'échantillon réalisées, n_{bd}, sont égales à des quantités fixes connues d'avance, comme il est décrit à la section 2.3.

Les estimations de ${}_{bd}t_r$, désignées par ${}_{bd}\hat{t}_{r,\text{greg}}$, sont obtenues au moyen de l'estimateur GREG *modifié* (Rao 2003, page 20) donné par :

$$ {}_{bd}\hat{t}_{r,\text{greg}} = \sum_{k \in s} {}_{bd}w_k\; y_{r,k} \quad (2.2.3) $$

où

$$ {}_{bd}w_k = a_k\, {}_{bd}\delta_k + ({}_{bd}t_{\mathbf{x}} - {}_{bd}\hat{t}_{\mathbf{x},ht})' \left[\sum_{k \in s} (a_k \mathbf{x}_k \mathbf{x}_k'/c_k)\right]^{-1} a_k\, \mathbf{x}_k/c_k $$

désigne les poids d'échantillonnage, $\mathbf{x}_k$ indique un vecteur de valeurs des variables auxiliaires, c_k est une constante connue, avec ${}_{bd}t_{\mathbf{x}} = \sum_{k \in U_{bd}} \mathbf{x}_k$ et ${}_{bd}\hat{t}_{\mathbf{x},ht} = \sum_{k \in s_{bd}} \mathbf{x}_k a_k$. L'estimateur (2.2.3) peut être calculé sous le modèle de superpopulation *de travail* suivant

$$ y_{r,k} = \mathbf{x}_k' \boldsymbol{\beta}_r + \varepsilon_{r,k} \quad (2.2.4) $$

où $\boldsymbol{\beta}_r$ est un vecteur inconnu des paramètres de régression fixes et $\varepsilon_{r,k}$ est le résidu aléatoire. L'espérance du modèle, E_m, et la variance du modèle, V_m, sont données respectivement par $E_m({}_r\varepsilon_k) = 0$; $V_m(\varepsilon_{r,k}) = c_k \sigma_r^2$; $E_m(\varepsilon_{r,k}, \varepsilon_{r,i}) = 0$ si $k \neq i$.

La variance d'échantillonnage approximative de l'estimateur GREG modifié sous échantillonnage équilibré est :

$$ V_p({}_{bd}\hat{t}_{r,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) = \frac{N}{N - Q} \sum_{k \in U} \left(\frac{1}{\pi_k} - 1\right) {}_{bd}\eta_{r,k}^2, \quad (2.2.5) $$

avec

$$ {}_{bd}\eta_{r,k} = \begin{cases} \varepsilon_{r,k} - \mathbf{z}_k'\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} & \text{pour } k \in U_{bd} \\ -\mathbf{z}_k'\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} & \text{pour } k \in U_{bd\bar{}} \end{cases}, \quad (2.2.6) $$

où

$$ {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon} = \left[\sum_{k \in U} \mathbf{z}_k \mathbf{z}_k' (1/\pi_k - 1)\right]^{-1} \sum_{k \in U} \mathbf{z}_k\, \varepsilon_{r,k}\, {}_{bd}\delta_k (1/\pi_k - 1), $$

avec $U_{bd\bar{}}$ le sous-ensemble de U complémentaire de U_{bd}. Une preuve de (2.2.5) est donnée à la section 2.5.

Les probabilités d'inclusion, π_k, et les tailles d'échantillon de domaine, n_{bd}, sont déterminées par une méthode visant à minimiser la taille globale d'échantillon, n, tout en garantissant que les variances d'échantillonnage seront inférieures aux seuils de niveau de précision préétablis, ${}_{bd}\bar{V}_r : V_p({}_{bd}\hat{t}_{r,\text{greg}} \,|\, \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\bar{V}_r$ $(b = 1, ..., B\,;\, d = 1, ..., M_b\,;\, r = 1, ..., R)$. Les détails techniques sont décrits à la section 3.

Il convient de souligner que deux ensembles différents de covariables ont été introduits afin de mettre en relief le fait que l'ensemble de covariables disponibles à l'étape de l'élaboration du plan (variables $\mathbf{z}$) pourrait être différent de l'ensemble disponible à l'étape de l'estimation (variables $\mathbf{x}$), même si, dans de nombreuses situations pratiques, ils

peuvent être identiques. Par exemple, les covariables au stade de l'estimation pourraient être mises à jour en fonction de celles disponibles à l'étape de l'élaboration du plan. Dans notre contexte (voir la section 2.3), les vecteurs $\mathbf{z}_k$ sont caractérisés comme le spécifie l'expression (2.3.2), étant définis uniquement par les variables indicatrices d'appartenance au domaine et par les probabilités d'inclusion, tandis que les vecteurs $\mathbf{x}_k$ pourraient contenir les valeurs de certaines autres variables plus explicatives du phénomène d'intérêt. Par exemple, dans le contexte des enquêtes-entreprises, les variables $\mathbf{x}$ pourraient inclure, entre autres, le nombre d'employés et le chiffre d'affaires.

2.3 Échantillonnage équilibré pour la stratification marginale

Les plans de sondage à stratification multidimensionnelle peuvent être traités dans le contexte de l'*échantillonnage équilibré*.

La définition d'un échantillon équilibré dépend du cadre inférenciel adopté. Dans l'approche fondée sur un modèle, un échantillon est dit *équilibré* sur un ensemble de variables auxiliaires s'il y a égalité entre les moyennes d'échantillon et les moyennes de population connues des variables auxiliaires (Valliant, Dorfman et Royall 2000). Dans le cas de l'approche fondée sur le plan de sondage (ou assistée par modèle) considérée ici, un échantillon est *équilibré* quand les estimations d'Horvitz-Thompson des totaux des variables auxiliaires sont égales aux totaux de population connus (Deville et Tillé 2004).

Afin de définir l'échantillonnage équilibré sous l'approche fondée sur le plan ou assistée par modèle, introduisons la définition générale du plan d'échantillonnage en tant qu'une loi de probabilité $p(\cdot)$ sur l'ensemble S de tous les sous-ensembles s de la population U, telle que $\sum_{s\in\mathbf{S}} p(s) = 1$, où $p(s)$ est la probabilité que l'échantillon s soit tiré. Chaque ensemble s peut être représenté par le résultat $\boldsymbol{\lambda}' = (\lambda_1, ..., \lambda_k, ..., \lambda_N)$ d'un vecteur de N variables aléatoires. Soit $\boldsymbol{\pi}' = (\pi_1, ..., \pi_k, ..., \pi_N)$ le vecteur de probabilités d'inclusion, où $\boldsymbol{\pi} = E_p(\boldsymbol{\lambda}) = \sum_{s\in\mathbf{S}} p(s)\boldsymbol{\lambda}$, avec $E_p(\cdot)$ la valeur espérée sous échantillonnage répété. Soit $\mathbf{z}'_k = (z_{1k}, ..., z_{hk}, ..., z_{Qk})$ un vecteur de Q variables auxiliaires disponible pour chaque unité de la population. Le plan d'échantillonnage $p(s)$ avec probabilités d'inclusion $\boldsymbol{\pi}$ est dit *équilibré* par rapport aux Q variables auxiliaires si, et uniquement si, il satisfait les équations d'équilibrage données par (2.2.2) pour tout $s \in S$ tel que $p(s) > 0$.

Supposons qu'un vecteur de probabilités d'inclusion $\boldsymbol{\pi}$ compatible avec les distributions d'échantillonnage marginales n_{bd} $(b = 1, ..., B\ ;\ d = 1, ..., M_b)$ est disponible, autrement dit que

$$n_{bd} = \sum_{k\in U_{bd}} \pi_k (b = 1, ..., B\ ;\ d = 1, ..., M_b). \quad (2.3.1)$$

Le plan à stratification multidimensionnelle représente un cas particulier de plan équilibré où, pour l'unité k, le vecteur de variables auxiliaires est donné par

$$\mathbf{z}'_k = (\overbrace{0, ..., \pi_k, ..., 0}^{b=1}, ..., \overbrace{0, ..., \pi_k, ..., 0}^{b=B})$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{Q}$$
$$= \pi_k({}_{11}\delta_k, ..., {}_{bd}\delta_k, ..., {}_{BM_B}\delta_k). \quad (2.3.2)$$

L'expression (2.3.2) définit les $\mathbf{z}_k$ comme étant des vecteurs contenant $(Q - B)$ zéros et B entrées égales à π_k aux emplacements indiquant les domaines auxquels appartient l'unité k. Quand le vecteur $\mathbf{z}_k$ est défini sous la forme (2.3.2), si la condition (2.3.1) tient, la sélection de l'échantillon qui satisfait le système d'*équations d'équilibrage* (2.2.2), $\sum_{k\in U} (\mathbf{z}_k\lambda_k)/\pi_k = \sum_{k\in s} \mathbf{z}_k$, garantit que les valeurs n_{bd} sont des quantités non aléatoires. Le premier membre des équations d'équilibrage (2.2.2) est $\sum_{k\in U} (\pi_k\ {}_{bd}\delta_k\lambda_k)/\pi_k = \sum_{k\in U_{bd}} \lambda_k = n_{bd}$, tandis que le deuxième membre est $\sum_{k\in U} \pi_k\ {}_{bd}\delta_k = \sum_{k\in U_{bd}} \pi_k = n_{bd}$.

Deville et Tillé (2004) ont proposé la *méthode du cube* qui permet de sélectionner des échantillons équilibrés (ou approximativement équilibrés) pour un grand ensemble de variables auxiliaires et par rapport à divers vecteurs de probabilités d'inclusion. En particulier, Deville et Tillé (2000) montrent qu'avec la spécification (2.3.2) des vecteurs $\mathbf{z}_k$, les équations d'équilibrage (2.2.2) peuvent être satisfaites exactement. La méthode du cube est implémentée par un algorithme amélioré pour grands ensembles de données (Chauvet et Tillé 2006) disponible dans un code de logiciel gratuit qui peut être téléchargé à partir du site Web *http://www.insee.fr/fr/nom_df_met/outils_stat/cube/accueil_cube.htm*.

2.4 Estimateur GREG direct modifié

En nous inspirant de Lehtonen et coll. (2003), nous pouvons exprimer l'estimateur (2.2.3) sous la forme générale

$${}_{bd}\hat{t}_{r,\,\mathrm{greg}} = \sum_{k\in U_{bd}} \tilde{y}_{r,k} + \sum_{k\in s_{bd}} a_k\,(y_{r,k} - \tilde{y}_{r,k}) \quad (2.4.1)$$

où $\tilde{y}_{r,k}$ dénote la prédiction de $y_{r,k}$ sous le modèle de superpopulation hypothétique. Les prédictions $\{\tilde{y}_{r,k}\ ;\ k \in U\}$ diffèrent d'une spécification de modèle à l'autre, selon la forme fonctionnelle et selon le choix des variables auxiliaires. L'estimateur (2.2.3) est calculé sous le modèle de superpopulation de travail (2.2.4). Les prédictions $\tilde{y}_{r,k}$ sont alors obtenues par

$$\tilde{y}_{r,k} = \mathbf{x}'_k\,\hat{\boldsymbol{\beta}}_r, \quad (2.4.2)$$

où

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{k\in s} \mathbf{x}_k\mathbf{x}'_k\,a_k/c_k\right)^{-1} \sum_{k\in s} \mathbf{x}_k\,y_{r,k}\,a_k/c_k. \quad (2.4.3)$$

Soulignons que le modèle linéaire (2.2.4) permet de définir l'estimateur en ne connaissant que les totaux de domaine de l'information auxiliaire et les valeurs $\mathbf{x}_k$ pour les unités d'échantillonnage. Toutefois, connaissant les valeurs $\mathbf{x}_k$ pour chaque unité $k \in U$, il est possible de construire des estimateurs donnant des prédictions plus efficaces $\tilde{y}_{r,k}$ obtenues par des modèles linéaires généralisés (Lehtonen et Veijanen 1998) ou des méthodes de régression non paramétriques (Montanari et Ranalli 2003).

Comme l'a souligné Rao (2003, page 20), l'estimateur (2.2.3) est approximativement sans biais sous le plan. Quand la taille globale d'échantillon augmente, même si la taille d'échantillon de domaine n_{bd} est faible. En outre, la somme des estimations ${}_{bd}\hat{t}_{r,\text{greg}}$ sur l'ensemble des domaines d'une partition est calée sur l'estimation GREG habituelle du total, $\sum_{d=1}^{M_b} {}_{bd}\hat{t}_{r,\text{greg}} = \sum_{k\in s} y_{r,k}\, a_k[1 + (\sum_{k\in U} \mathbf{x}_k - \sum_{k\in s} \mathbf{x}_k a_k)' (\sum_{k\in s} \mathbf{x}_k \mathbf{x}_k' a_k / c_k)^{-1} \mathbf{x}_k / c_k]$.

2.5 Variances d'échantillonnage

Afin d'établir l'expression de la variance (2.2.5), considérons les résultats donnés par Deville et Tillé (2005). Ces auteurs ont proposé d'approximer la variance de l'estimateur d'Horvitz-Thompson $\hat{t}_{r,ht} = \sum_{k\in s} y_{r,k}\, a_k$ du total $t_r = \sum_{k\in U} y_{r,k}$ en supposant que l'échantillonnage équilibré peut être considéré comme un échantillonnage de Poisson conditionnel et en supposant que, du moins pour les grandes tailles d'échantillon, les probabilités d'inclusion π_k sont de bonnes approximations des probabilités d'inclusion du plan de Poisson. Posant que, par la voie de l'échantillonnage de Poisson, le vecteur $(\hat{t}_{r,ht},\ \hat{t}'_{\mathbf{z},ht})'$ suit approximativement une loi multinormale, ils proposent une bonne approximation de la variance d'échantillonnage donnée par

$$\begin{aligned} V_p(\hat{t}_{r,ht}|\hat{t}_{\mathbf{z},ht}=t_\mathbf{z}) &= V_p(\hat{t}_{r,ht} + (t_\mathbf{z} - \hat{t}_{\mathbf{z},ht})'\, \mathbf{B}_{\mathbf{z},y}) \\ &= V_p(\hat{t}_{r,ht} - \hat{t}'_{\mathbf{z},ht}\, \mathbf{B}_{\mathbf{z},y}) \\ &= V_p\left(\sum_{k\in s} a_k (y_{r,k} - \mathbf{z}'_k\, \mathbf{B}_{\mathbf{z},y})\right) \\ &\cong \frac{N}{N-Q} \sum_{k\in U} \left(\frac{1}{\pi_k} - 1\right)(y_{r,k} - \mathbf{z}'_k\, \mathbf{B}_{\mathbf{z},y})^2 \quad (2.5.1) \end{aligned}$$

où $\mathbf{B}_{\mathbf{z},y} = [\sum_{k\in U} \mathbf{z}_k \mathbf{z}'_k (1/\pi_k - 1)]^{-1} \sum_{k\in U} \mathbf{z}_k\, y_{r,k} (1/\pi_k - 1)$. L'expression (2.5.1) a été validée par un ensemble de simulations.

Considérons maintenant l'approximation linéaire, ${}_{bd}\hat{t}^*_{r,\text{greg}}$, de l'estimateur GREG, dont le calcul peut être obtenu conformément à Särndal, Swensson et Wretman (1992, pages 450-451) par

$$\begin{aligned} {}_{bd}\hat{t}_{r,\text{greg}} \cong {}_{bd}\hat{t}^*_{r,\text{greg}} &= \sum_{k\in U_{bd}} \mathbf{x}'_k \boldsymbol{\beta}_r + \sum_{k\in s_{bd}} a_k\, \varepsilon_{r,k} \\ &= \sum_{k\in U_{bd}} \mathbf{x}'_k \boldsymbol{\beta}_r + \sum_{k\in s} a_k\, \varepsilon_{r,k}\, {}_{bd}\delta_k. \quad (2.5.2) \end{aligned}$$

Partant des expressions (2.5.1) et (2.5.2), il est possible de dériver le résultat suivant

$$\begin{aligned} V_p({}_{bd}\hat{t}_{r,\text{greg}}|\hat{t}_{\mathbf{z},ht} = t_\mathbf{z}) &\cong V_p({}_{bd}\hat{t}^*_{r,\text{greg}}|\hat{t}_{\mathbf{z},ht} = t_\mathbf{z}) \\ &= V_p\left(\sum_{k\in s} a_k\, \varepsilon_{r,k}\, {}_{bd}\delta_k | \hat{t}_{\mathbf{z},ht} = t_\mathbf{z}\right) \\ &= V_p\left(\sum_{k\in s} a_k \varepsilon_{r,k}\, {}_{bd}\delta_k + (t_\mathbf{z} - \hat{t}_{\mathbf{z},ht})'\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\right) \\ &= V_p\left(\sum_{k\in s} a_k\, (\varepsilon_{r,k}\, {}_{bd}\delta_k - \mathbf{z}'_k\, {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\right) \\ &= V_p\left(\sum_{k\in s} a_k\, {}_{bd}\eta_{r,k}\right) \\ &\cong \frac{N}{N-Q} \sum_{k\in U} \left(\frac{1}{\pi_k} - 1\right) {}_{bd}\eta^2_{r,k}, \end{aligned}$$

où ${}_{bd}\eta^2_{r,k}$ est défini dans (2.2.6).

La variance d'échantillonnage approximative de ${}_{bd}\hat{t}_{r,\text{greg}}$ dépend des résidus de l'ensemble complet d'unités, à cause de la sélection équilibrée. Par conséquent, les unités qui n'appartiennent pas à U_{bd} ont une influence sur la variance d'échantillonnage de l'estimateur.

Examinons maintenant le cas unidomaine univarié et supposons que le sondage ne comporte qu'un seul paramètre cible, ${}_{bd}t_r$. En outre, supposons que l'échantillon sélectionné satisfait les *équations d'équilibrage*, $\hat{t}_{\mathbf{z},ht} = t_\mathbf{z}$, quand la taille globale d'échantillon n est fixe.

En suivant les arguments proposés par Särndal et coll. (1992 ; résultat 12.2.1, page 452), il est facile de prouver que, dans ce contexte d'échantillonnage, chaque unité k pourrait être sélectionnée avec $(Q \times R)$ probabilités d'inclussion optimales différentes, ${}_{bd}\ddot{\pi}_{r,k}$ ($b = 1, ..., B$; $d = 1, ..., M_b$; $r = 1, ..., R$)

$${}_{bd}\ddot{\pi}_{r,k} = n\, |\, {}_{bd}\eta_{r,k}\, | \Big/ \sum_{i\in U} |\, {}_{bd}\eta_{r,i}\, |,$$

ce qui permet d'atteindre les $(Q \times R)$ bornes inférieures différentes, ${}_{bd}V^*_{r|n}$, des variances approximatives :

$$V_p({}_{bd}\hat{t}_{r,\text{greg}}|\hat{t}_{\mathbf{z},ht} = t_\mathbf{z}) \geq {}_{bd}V^*_{r|n} = \frac{N}{N-Q}\left[\frac{1}{n}\left(\sum_{k\in U} |\, {}_{bd}\eta_{r,k}\, |\right)^2 - \sum_{k\in U} {}_{bd}\eta^2_{r,k}\right].$$

Enfin, mentionnons que Tillé et Favre (2005) donnent un critère pour l'obtention d'une prédiction ${}_{bd}\hat{\eta}_{r,k}$ des valeurs ${}_{bd}\eta_{r,k}$ qui peut être utilisée dans les contextes d'échantillonnage répété.

3. Algorithmes d'échantillonnage pour la détermination des tailles d'échantillon de domaine

Les probabilités d'inclusion π_k et les tailles d'échantillon de domaine dérivées, $n_{bd} = \sum_{k \in U_{bd}} \pi_k$, sont obtenues par une méthode en deux étapes : i) à la première étape, appelée *optimisation*, les probabilités d'inclusion préliminaires, π'_k, sont déterminées en résolvant un problème de minimisation sous contraintes et ii) à la deuxième étape, appelée calage, les probabilités d'inclusion, π_k, sont obtenues sous forme d'une légère modification des π'_k ; le *calage* est mis en œuvre pour s'assurer que les tailles d'échantillon de domaine n_{bd} soient des nombres entiers.

Comme nous l'illustrons ci-après, les valeurs π_k peuvent être exprimées sous forme de fonctions implicites des résidus inconnus ${}_{bd}\eta^2_{r,k}$. Toutefois, dans le contexte des sondages réels, la détermination des probabilités d'inclusion π_k peut se faire en utilisant les prédictions ${}_{bd}\hat{\eta}^2_{r,k}$ au lieu de ${}_{bd}\eta^2_{r,k}$. Il s'agit d'un problème général concernant la planification des plans d'échantillonnage, parce que les variances sont habituellement des quantités inconnues qui peuvent être estimées convenablement. Dans le contexte des sondages répétés, l'effet de l'utilisation des estimations ${}_{bd}\hat{\eta}^2_{r,k}$ pour remplacer ${}_{bd}\eta^2_{r,k}$ peut être évalué en calculant les variances d'échantillonnage après la collecte des données. Les résultats empiriques peuvent alors être utilisés pour introduire les corrections appropriées dans la planification du plan de sondage suivant. Toutefois, comme l'illustre l'analyse empirique, ainsi que Falorsi et Righi (2008), la stratégie proposée semble être efficace et suffisamment robuste aux petits écarts par rapport aux conditions idéales.

Les sections 3.1 et 3.2 décrivent, respectivement, les deux étapes de l'algorithme de détermination des tailles d'échantillon de domaine. Une *règle de répartition simplifiée*, qui semble avoir du mérite dans de nombreux contextes de sondage réel, est décrite à la section 3.3.

3.1 Première étape : optimisation

Les probabilités d'inclusion π'_k peuvent être définies comme étant la solution du problème ci-après de programmation non linéaire à N inconnues, π'_k, sous les contraintes $(N + Q \times R)$

$$\begin{cases} \text{Min}\left(\sum_{k \in U} \pi'_k\right) \\ \\ \dfrac{N}{N-Q}\sum_{k \in U}\left(\dfrac{1}{\pi'_k} - 1\right){}_{bd}\eta^2_{r,k} \le {}_{bd}\bar{V}_r \\ \quad (b = 1, ..., B\ ; d = 1, ..., M_b\ ; r = 1, ..., R) \\ \\ 0 < \pi'_k \le 1 \quad (k = 1, ..., N). \end{cases} \tag{3.1.1}$$

Une solution numérique de (3.1.1) peut être obtenue en considérant les algorithmes élaborés pour la répartition multivariée dans les sondages stratifiés. Ces algorithmes permettent de trouver les valeurs inconnues $v_h > 0$ $(h = 1, 2, ...)$ qui représentent la solution du problème non linéaire $\text{Min}(\sum_h \nu_h)$ sous la contrainte $\sum_h A_{rh}/\nu_h \le \bar{A}_r$, où A_{rh} et $\bar{A}_r$ $(r = 1, 2, ...)$ sont des quantités positives connues.

Bethel (1989) fait appel au théorème de Kuhn-Tucker pour montrer qu'il existe une solution au problème susmentionné. Il décrit un algorithme simple et discute de ses propriétés de convergence. Chromy (1987) élabore un algorithme qui convient pour les tableurs automatisés, mais sans donner de preuve explicite qu'il est toujours convergent. Nous décrivons ci-après une légère modification de l'algorithme de Chromy qui permet de résoudre le problème (3.1.1) en garantissant que les inégalités $0 < \pi'_k \le 1$ $(k = 1, ..., N)$ sont respectées. Après l'initialisation, l'algorithme trouve les valeurs de π'_k par itération des deux étapes *Calcul* et *Vérification*. En ce qui concerne le problème de convergence, il convient de souligner que l'algorithme de Chromy a été utilisé principalement pour les plans d'échantillonnage stratifiés et, en effet, la documentation fait référence à des échantillons stratifiés. La littérature sur l'échantillonnage appliqué offre un grand nombre de preuves empiriques de l'utilisation fructueuse de l'algorithme dans ce contexte. Mentionnons que, dans la modification de l'algorithme de Chromy proposée ici, les unités d'échantillonnage sont traitées comme des strates et la répartition résultante, qui est fractionnaire, définit les probabilités d'inclusion. En outre, il n'existe pas de preuve formelle que l'algorithme modifié proposé converge dans ces conditions. Néanmoins, dans toutes les expériences empiriques que nous avons élaborées, il a convergé systématiquement et nous n'avons rencontré aucune condition critique.

Initialisation : à l'itération initiale $(\tau = 0)$, poser ${}^{\tau}\gamma_k = 1$ $(k = 1, ..., N)$.

Calcul : l'itération générique $(\tau = 1, 2, ...)$ consiste en une série d'étapes désignées par $u = (0, 1, 2, ...)$.

- À l'étape initiale $(u = 0)$, poser ${}^{\tau,u}_{bd}\phi_r = 1$ et calculer

$${}^{\tau}_{bd}V_{0r} = \frac{N}{N-Q}\sum_{k \in U} {}_{bd}\eta^2_{r,k}\ {}^{\tau}\gamma_k.$$

- Aux étapes subséquentes $(u = 1, 2, ...)$, calculer les valeurs des équations suivantes

$${}^{\tau,u}\pi_k = \left[(1 - {}^{\tau}\gamma_k) + {}^{\tau}\gamma_k \frac{N}{N-Q}\sum_{b=1}^{B}\sum_{d=1}^{M_b}\sum_{r=1}^{R} {}^{\tau,u}_{bd}\phi_r\ {}_{bd}\eta^2_{r,k}\right]^{1/2}. \tag{3.1.2}$$

$$ {}^{\tau,u}_{bd}V_r = \frac{N}{N-Q}\sum_{k\in U}\frac{1}{{}^{\tau,u}\pi_k}\,{}_{bd}\eta^2_{r,k}\,{}^{\tau}\gamma_k, $$

et (3.1.3)

$$ {}^{\tau,u}_{bd}V'_r = {}^{\tau,u}_{bd}V_r + {}_{bd}{}^{\tau}V_{0r}. $$

- Si les deux conditions suivantes :

$$ {}^{\tau,u}_{bd}V'_r \le {}_{bd}\bar{V}_r \text{ et } {}^{\tau,u}_{bd}\phi_r({}^{\tau,u}_{bd}V'_r - {}_{bd}\bar{V}_r) = 0, \qquad (3.1.4) $$

sont respectées (pour tout $b = 1, ..., B$; $d = 1, ..., M_b$; $r = 1, ..., R$), alors l'étape Calcul s'arrête et les probabilités d'inclusion ${}^{\tau}\pi_k$ sont celles calculées dans l'équation (3.1.2). Sinon, les quantités mises à jour ${}^{\tau,u+1}_{bd}\phi_r$ sont calculées selon

$$ {}^{\tau,u+1}_{bd}\phi_r = {}^{\tau,u}_{bd}\phi_r[{}^{\tau,u}_{bd}V_r/({}^{\tau,u}_{bd}V'_r - {}^{\tau}_{bd}\bar{V}_r)]^2 \qquad (3.1.5) $$

et les équations (3.1.2) et (3.1.3) sont calculées à $u + 1$, répétitivement avec ${}^{\tau,u+1}_{bd}\phi_r$ remplaçant ${}^{\tau,u}_{bd}\phi_r$, jusqu'à ce que les conditions (3.1.4) soient respectées.

Vérification : si la condition ${}^{\tau}\pi_k \le 1$ est vraie pour tout k, alors l'algorithme s'arrête et les valeurs de π'_k sont fixées à $\pi'_k = {}^{\tau}\pi_k$. Sinon, les valeurs de ${}^{\tau}\gamma_k$ sont mises à jour comme étant ${}^{\tau+1}\gamma_k = 1$ si ${}^{\tau}\pi_k \le 1$ et ${}^{\tau+1}\gamma_k = 0$ si ${}^{\tau}\pi_k > 1$. Le calcul est itéré à $\tau + 1$ avec ${}^{\tau+1}\gamma_k$ remplaçant ${}^{\tau}\gamma_k$. Nous avons développé une macro SAS permettant de résoudre le problème (3.1.1) qui peut être obtenue sur demande.

3.2 Deuxième étape : calage

Pour commencer, les quantités n_{bd} sont définies en arrondissant les résultats des Q sommes, $\sum_{k\in U_{bd}}\pi'_k (b = 1, ..., B\,;\, d = 1, ..., M_b)$. Parfois, une manipulation supplémentaire des données peut être nécessaire pour s'assurer que la condition $\sum_{d=1}^{M_b} n_{bd} = n$ est respectée pour chaque b. Ensuite, les probabilités π_k sont obtenues en tant que solution du *problème de calage*

$$ \begin{cases} \text{Min}\left(\sum_{k\in U} G(\pi_k;\pi'_k)\right) \\ \sum_{k\in U}\pi_k = n,\ \sum_{k\in U_{bd}}\pi_k = n_{bd} \\ \qquad (b = 1, ..., B\,;\, d = 1, ..., M_b - 1), \end{cases} \qquad (3.2.1) $$

où $G(\pi_k\,;\,\pi'_k)$ est une fonction de distance entre π_k et π'_k. Notons que le problème (3.2.1) peut être résolu par l'algorithme bien connu d'*ajustement proportionnel itératif* (Bishop, Fienberg et Holland 1975) ou l'algorithme d'*ajustement proportionnel itératif généralisé* (APIG ; Dykstra et Wollan 1987). La fonction de distance logarithmique $G(\pi_k\,;\,\pi'_k) = \pi_k \ln(\pi_k/\pi'_k) - (\pi_k + \pi'_k)$ évite de devoir définir des probabilités π_k inférieures à zéro, tandis que l'APIG évite d'obtenir des valeurs de π_k supérieures à 1.

3.3 Une règle de répartition simplifiée

Dans de nombreux contextes de sondage réel où la taille globale d'échantillon n est fixe et où il n'existe pas suffisamment d'information pour obtenir de bonnes prédictions ${}_{bd}\hat{\eta}^2_{r,k}$ des valeurs ${}_{bd}\eta^2_{r,k}$, il est possible d'appliquer la procédure qui suit. Premièrement, les tailles d'échantillon marginales n_{bd} sont déterminées suivant une règle assez simple

$$ n_{bd} = \alpha_b\, n(N_{bd}/N) + (1-\alpha_b)n/M_b, \qquad (3.3.1) $$

où $\alpha_b (0 \le \alpha_b \le 1)$ est une constante fixée qui a été déterminée correctement. La règle (3.3.1) s'avère être un compromis entre la répartition proportionnelle à la taille de la population ($\alpha_b = 1$) et la répartition uniforme pour chaque domaine d'une partition donnée ($\alpha_b = 0$).

Les probabilités π_k sont alors obtenues comme solution du problème de calage (3.2.1) où les tailles d'échantillon marginales sont calculées comme il est indiqué plus haut et les probabilités initiales π'_k sont fixées uniformément et égales à $\pi'_k = n/N$. Les probabilités d'inclusion résultantes ne sont plus *optimales*, au sens décrit plus haut et ne garantissent pas que les variances d'échantillonnage sont inférieures au seuil préétabli de niveau de précision. Cependant, elles sont calculées selon une méthode raisonnable qui peut être mise en œuvre équitablement et qui représente donc un point de référence intéressant en ce qui concerne tout contexte de sondage réel.

4. La variance anticipée

Un critère souvent utilisé pour planifier les stratégies d'échantillonnage est le contrôle de la variance anticipée, qui peut être définie comme étant :

$$ \text{AV}({}_{bd}\hat{t}_{r,\text{greg}}\,|\,\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) = E_m E_p({}_{bd}\hat{t}_{r,\text{greg}} - {}_{bd}t_r\,|\,\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}})^2. \quad (4.1) $$

Le résultat qui suit peut être obtenu sous les hypothèses du modèle (2.2.4) en utilisant les résultats donnés à la section (2.5) :

$$ \begin{aligned} &\text{AV}({}_{bd}\hat{t}_{r,\text{greg}}\,|\,\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) \\ &\quad \cong E_m V_p({}_{bd}\hat{t}^*_{r,\text{greg}}\,|\,\hat{\mathbf{t}}_{\mathbf{z},ht} = \mathbf{t}_{\mathbf{z}}) \\ &\quad = E_m\left[\frac{N}{N-Q}\sum_{k\in U_{bd}}\left(\frac{1}{\pi_k}-1\right)\varepsilon^2_{r,k}\right. \\ &\qquad \frac{N}{N-Q}\sum_{k\in U}\left(\frac{1}{\pi_k}-1\right)(\mathbf{z}'_k\,{}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 \\ &\qquad \left. -2\,\frac{N}{N-Q}\sum_{k\in U_{bd}}\left(\frac{1}{\pi_k}-1\right)\varepsilon_{r,k}\,\mathbf{z}'_k\,{}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}\right] \\ &\quad = \frac{N}{N-Q}\sum_{k\in U}\left(\frac{1}{\pi_k}-1\right){}^{a}_{bd}\eta^2_{r,k} \end{aligned} \qquad (4.2) $$

avec

$$ {}_{bd}^{a}\eta_{r,k}^{2} = \begin{cases} \sigma_r^2 c_k (1 - g_{kk})^2 + \sigma_r^2 \sum_{j(\neq k)\in U_{bd}} g_{kj}^2 c_j & \text{si } k \in U_{bd} \\ \sigma_r^2 \sum_{j\in U_{bd}} g_{kj}^2 c_k & \text{autrement,} \end{cases} $$

où $(g_{k1}, ..., g_{kj}, ..., g_{kN}) = \mathbf{g}_k' = \mathbf{z}_k'(\mathbf{Z}_U' \boldsymbol{\Omega}_U^{-1} \mathbf{Z}_U)^{-1}\mathbf{Z}_U' \boldsymbol{\Omega}_U^{-1}$, $\mathbf{Z}_U = \text{col}\{\mathbf{z}_k'\}_{k=1}^{N}$, $\boldsymbol{\Omega}_U^{-1} = \text{diag}\{1/\pi_k - 1\}_{k=1}^{N}$. L'expression (4.2) a été obtenue en utilisant les deux résultats suivants :

$$ \begin{aligned} & E_m(\mathbf{z}_k' {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon})^2 \\ & = \sigma_r^2 \mathbf{z}_k'(\mathbf{Z}_U'\boldsymbol{\Omega}_U^{-1}\mathbf{Z}_U)^{-1}\mathbf{Z}_U'\boldsymbol{\Omega}_U^{-1} {}_{bd}\mathbf{V}_r \boldsymbol{\Omega}_U^{-1}\mathbf{Z}_U(\mathbf{Z}_U'\boldsymbol{\Omega}_U^{-1}\mathbf{Z}_U)^{-1}\mathbf{z}_k \\ & = \sigma_r^2 \mathbf{g}_k' {}_{bd}V_r \mathbf{g}_k = \sigma_r^2 \sum_{j\in U_{bd}} g_{kj}^2 c_k, \end{aligned} $$

$$ \begin{aligned} & E_m(\varepsilon_{r,k} \mathbf{z}_k' {}_{bd}\mathbf{B}_{\mathbf{z},\varepsilon}) = \mathbf{z}_k'(\mathbf{Z}_U'\boldsymbol{\Omega}_U^{-1}\mathbf{Z}_U)^{-1}\mathbf{Z}_U'\boldsymbol{\Omega}_U^{-1} E_m \\ & (\varepsilon_{r,k}\mathbf{I}_N(\varepsilon_{r,1} {}_{bd}\delta_1, ..., \varepsilon_{r,k\,bd}\delta_k, ..., \varepsilon_{r,N} {}_{bd}\delta_N)') \\ & = \sigma_r^2 g_{kk} c_k {}_{bd}\delta_k, \end{aligned} $$

où ${}_{bd}\mathbf{V}_r = \text{diag}\{c_k {}_{bd}\delta_k\}_{k=1}^{N}$ et $\mathbf{I}_N = \text{diag}\{1\}_{k=1}^{N}$.

Le résultat (4.2) montre qu'il est possible de définir une stratégie d'échantillonnage dont le but est de contrôler les variances anticipées. En effet, si les quantités ${}_{bd}^{a}\eta_{r,k}^2$ (ou leurs prédictions correctes ${}_{bd}^{a}\hat{\eta}_{r,k}^2$) sont utilisées pour remplacer les résidus ${}_{bd}\eta_{r,k}^2$, le problème (3.1.1) définit un plan d'échantillonnage qui garantit que les conditions $\text{AV}({}_{bd}\hat{t}_{r,\text{greg}} | \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\overline{V}_r$ ($b = 1, ..., B$; $d = 1, ..., M_b$; $r = 1, ..., R$) seront respectées.

Le résultat qui suit est intéressant. Dans le cas particulier d'une partition unique, si les probabilités d'inclusion, π_k, et les facteurs d'hétéroscédasticité, c_k, sont relativement constants dans chaque domaine, la sélection d'un échantillon équilibré réduit la variance anticipée. Ce résultat est démontré dans Falorsi et Righi (2008).

5. Brève extension au cas d'un estimateur indirect simple de petit domaine

Si une partition donnée de la population produit un trop grand nombre de domaines, il se peut que des contraintes budgétaires obligent à définir une erreur d'échantillonnage préétablie trop grande des estimateurs directs des domaines de la partition ; le cas échéant, il est parfois nécessaire d'adopter un estimateur de petit domaine indirect afin de contrôler les erreurs quadratiques moyennes des estimations des domaines de la partition. Dans la suite de l'exposé, nous montrerons comment la stratégie d'échantillonnage décrite aux sections 2 et 3 peut être étendue au cas d'un estimateur indirect simple de petit domaine. Considérons le cas suffisamment général dans lequel le vecteur $\mathbf{x}_k$ des covariables auxiliaires contient une coordonnée à l'origine (constante), telle que $N_{bd} = \sum_{k\in s\,bd} w_k$.

Soit $\ddot{b}$ la partition pour laquelle il est nécessaire d'adopter un estimateur de petit domaine indirect et considérons le modèle (7.1.1) décrit dans Rao (2005, page 116). Dans le contexte étudié ici, le modèle pour l'estimateur direct, ${}_{\ddot{b}\ddot{d}}\hat{\bar{t}}_{r,\text{greg}} = {}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{greg}}/N_{\ddot{b}\ddot{d}}$, du domaine $\ddot{b}\ddot{d}$ peut être défini comme étant

$$ {}_{\ddot{b}\ddot{d}}\hat{\bar{t}}_{r,\text{greg}} = {}_{\ddot{b}\ddot{d}}\mathbf{a}'\boldsymbol{\varphi}_r + {}_{\ddot{b}\ddot{d}}h\ {}_{\ddot{b}\ddot{d}}v_r + {}_{\ddot{b}\ddot{d}}u_r \quad (d = 1, ..., M_{\ddot{b}}\ ;\ r = 1, ..., R) \qquad (5.1) $$

où ${}_{\ddot{b}\ddot{d}}\mathbf{a}$ est un vecteur de dimension $p \times 1$ des covariables au niveau du domaine, $\boldsymbol{\varphi}_r$ est un vecteur de dimension $p \times 1$ inconnu des coefficients de régression, ${}_{\ddot{b}\ddot{d}}h$ est une quantité connue reliée au domaine $\ddot{b}\ddot{d}^{\text{th}}$, ${}_{\ddot{b}\ddot{d}}v_r \sim \text{iid}(0, {}_{\ddot{b}}\sigma_{r\text{v}}^2)$ indépendante de l'erreur d'échantillonnage ${}_{\ddot{b}\ddot{d}}u_r \sim$ approximativement $\text{ind}(0, {}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2)$, avec ${}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 = V_p({}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{greg}} | \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}})/N_{\ddot{b}\ddot{d}}^2$. Pour les valeurs ${}_{\ddot{b}}\sigma_{r\text{v}}^2$ et ${}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2$ connues, l'estimateur BLUP de ${}_{\ddot{b}\ddot{d}}t_r$ est

$$ {}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{blup}} = N_{\ddot{b}\ddot{d}}({}_{\ddot{b}\ddot{d}}\gamma_r\ {}_{\ddot{b}\ddot{d}}\hat{\bar{t}}_{r,\text{greg}} + (1 - {}_{\ddot{b}\ddot{d}}\gamma_r)\ {}_{\ddot{b}\ddot{d}}\mathbf{a}'\hat{\boldsymbol{\varphi}}_r) \quad (5.2) $$

avec

$$ {}_{\ddot{b}\ddot{d}}\gamma_r = {}_{\ddot{b}}\sigma_{r\text{v}}^2\ {}_{\ddot{b}\ddot{d}}h^2/({}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 + {}_{\ddot{b}}\sigma_{r\text{v}}^2\ {}_{\ddot{b}\ddot{d}}h^2) \qquad (5.3) $$

et

$$ \hat{\boldsymbol{\varphi}} = \left[\sum_{d=1}^{M_{\ddot{b}}} {}_{\ddot{b}\ddot{d}}\mathbf{a}\ {}_{\ddot{b}\ddot{d}}\mathbf{a}'/({}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 + {}_{\ddot{b}}\sigma_{r\text{v}}^2\ {}_{\ddot{b}\ddot{d}}h^2)\right]^{-1} \left[\sum_{l=1}^{M_{\ddot{b}}} {}_{\ddot{b}l}\mathbf{a}\ {}_{\ddot{b}l}\hat{\bar{t}}_{r,\text{greg}}/({}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 + {}_{\ddot{b}}\sigma_{r\text{v}}^2\ {}_{\ddot{b}\ddot{d}}h^2)\right] \qquad (5.4) $$

L'EQM de l'estimateur BLUP est

$$ \text{EQM}({}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{blup}}) = N_{\ddot{b}\ddot{d}}^2 \Bigg[{}_{\ddot{b}\ddot{d}}\gamma_r\ {}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 + (1 - {}_{\ddot{b}\ddot{d}}\gamma_r)^2\ {}_{\ddot{b}\ddot{d}}\mathbf{a}'\left(\sum_{d=1}^{M_{\ddot{b}}} {}_{\ddot{b}\ddot{d}}\mathbf{a}\ {}_{\ddot{b}\ddot{d}}\mathbf{a}'/({}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2 + {}_{\ddot{b}}\sigma_{r\text{v}}^2\ {}_{\ddot{b}\ddot{d}}h^2)\right)^{-1} {}_{\ddot{b}\ddot{d}}\mathbf{a}\Bigg]. \qquad (5.5) $$

L'examen des expressions (5.3) et (5.5) montre que, pour une valeur donnée de la variance ${}_{\ddot{b}}\sigma_{r\text{v}}^2$, il est possible de contrôler l'$\text{EQM}({}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{blup}})$ dans la planification du plan d'échantillonnage en définissant une valeur appropriée de la variance ${}_{\ddot{b}\ddot{d}}\sigma_{r\bar{t}}^2$. La procédure itérative qui suit permet de trouver les probabilités d'inclusion π_k' qui garantissent que la taille d'échantillon est minimale et que les contraintes $V_p({}_{bd}\hat{t}_{r,\text{greg}} | \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\overline{V}_r$ (pour $b \neq \ddot{b}$; $d = 1, ..., M_b$; $r = 1, ..., R$) et $\text{EQM}({}_{\ddot{b}\ddot{d}}\hat{t}_{r,\text{blup}}) \leq {}_{\ddot{b}\ddot{d}}\overline{V}_r$ ($d = 1, ..., M_{\ddot{b}}$; $r = 1, ..., R$) sont respectées.

Initialisation : À l'itération initiale $(j = 0)$, trouver les probabilités d'inclusion ${}^{j}\pi'_k$, qui sont la solution du problème (3.1.1), en appliquant les contraintes $V_p({}_{\ddot{b}d}\hat{t}_{r,\text{greg}} \mid \hat{t}_{\mathbf{z},ht} = t_{\mathbf{z}}) \leq {}_{bd}\overline{V}_r$ (pour $b = 1,...,B$; $d = 1, ..., M_b$; $r = 1,...,R$).

Itération : L'itération générique $(j = 1, 2, ...)$ s'énonce comme il suit.

- Calculer ${}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}} = [N/(N^2_{\ddot{b}d}(N-Q))]\sum_{k\in U}[(1/{}^{j-1}\pi_k) - 1]_{\ddot{b}d}\eta^2_{r,k}$ $(d = 1, ..., M_{\ddot{b}}$; $r = 1, ..., R)$.
- Calculer ${}^{j}_{\ddot{b}d}\gamma_r$ et ${}^{j}\text{EQM}({}_{\ddot{b}d}\hat{t}_{r,\text{blup}})$ $(d = 1, ..., M_{\ddot{b}}$; $r = 1, ..., R)$ respectivement au moyen des équations (5.3) et (5.5) en utilisant les variances d'échantillonnage ${}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}}$ au lieu de ${}_{\ddot{b}d}\sigma^2_{r\bar{t}}$.
- Calculer ${}^{j}_{\ddot{b}d}\text{eff}_r = {}^{j}\text{EQM}({}_{\ddot{b}d}\hat{t}_{r,\text{blup}}) / ({}^{j}_{\ddot{b}d}\sigma^2_{r\bar{t}}\, N^2_{\ddot{b}d})$.
- Trouver les probabilités d'inclusion ${}^{j}\pi'_k$, qui sont la solution du problème (3.1.1), en utilisant les valeurs ${}^{j}_{\ddot{b}d}\eta^2_{r,k} = {}_{\ddot{b}d}\eta^2_{r,k}\ {}^{j}_{\ddot{b}d}\text{eff}_r$ $(d = 1, ..., M_{\ddot{b}}$; $r = 1, ..., R$; $k = 1, ..., N)$ pour remplacer les valeurs ${}_{\ddot{b}d}\eta^2_{r,k}$.

Vérification : Si la condition qui suit est satisfaite pour une petite quantité v, $\sum_{k\in U} |{}^{j-1}\pi_k - {}^{j}\pi_k| \leq v$, alors l'algorithme s'arrête et les probabilités d'inclusion π'_k sont celles calculées à l'itération j. Sinon, l'itération est calculée répétitivement jusqu'à ce que la condition soit respectée.

6. Analyse empirique

Afin de vérifier les propriétés empiriques de la stratégie d'échantillonnage proposée, nous avons effectué deux expériences qui ont toutes deux révélé que la stratégie proposée donne de bons résultats. La première expérience, qui portait sur des données artificielles, est décrite dans Falorsi et Righi (2008) ; la stratégie d'échantillonnage proposée à la section 2 est appliquée en entier, y compris la répartition de l'échantillon décrite aux sections 3.1 et 3.2. La deuxième expérience, une simulation basée sur des données réelles sur les entreprises, est décrite ci-après.

L'analyse porte sur la population de 1999 d'entreprises comptant de 1 à 99 employés de la division *Activités informatiques et activités rattachées* (classification à deux chiffres de la NACE Rév. 1). La base de données utilisée pour l'étude par simulation contient $N = 10\,392$ entreprises. Les variables d'intérêt choisies pour la simulation sont la *valeur ajoutée* et le *coût de la main-d'œuvre*. Les valeurs des variables, disponibles pour chaque unité de la population, proviennent d'une source de données administratives. Nous avons considéré deux partitions, à savoir la *région géographique* avec 20 domaines marginaux (DOM1) et le *groupe d'activités économiques* (classification à trois chiffres de la NACE Rév.1) avec six groupes distincts (DOM2) selon la *classe de taille* (définie en fonction du nombre de personnes employées : $1 = 1-4$; $2 = 5-9$; $3 = 10-19$; $4 = 20-99$) avec 24 domaines marginaux. Par conséquent, le nombre global de domaines marginaux est égal à 44, tandis que le nombre de strates par recoupement est de 480, mais seulement 360 de celles-ci possèdent une unité de population ou plus.

Dans la présente étude, n est fixé à 360. Nous avons étudié cinq plans d'échantillonnage, qui sont présentés au tableau 6.1. Les deux premiers, qui servent de référence, sont de simples plans à stratification unidimensionnelle avec échantillonnage aléatoire simple sans remise dans chaque strate. Le premier de ces plans, appelé STDOM1, est stratifié par la partition 1 et le deuxième, STDOM2, est stratifié par la partition 2. Pour STDOM1, les tailles marginales d'échantillon ont été définies par (3.3.1). Le paramètre α_1 et les tailles marginales d'échantillon connexes n_{1d} $(d = 1, ..., 20)$ assurent que le coefficient de variation (CV), exprimé en pourcentage, des estimations d'Horvitz-Thompson des totaux de la *variable auxiliaire Nombre d'employés* soient inférieurs à 34,5 % pour tous les domaines de la partition 1. De manière analogue, la valeur du paramètre α_2 a été définie au moyen de (3.3.1), ce qui assure que, sous le plan d'échantillonnage STDOM2, le CV, en pourcentage, des estimations d'Horvitz-Thompson des totaux de la variable auxiliaire est inférieur à 8,7 % pour tous les domaines de la partition 2. Dans la suite, nous désignons comme étant des *petits domaines* ceux dont la taille d'échantillon planifiée est plus grande que la taille d'échantillon calculée d'après une règle de répartition avec $\alpha_b = 1$ $(b = 1, 2)$. Ces petits domaines doivent être suréchantillonnés afin de pouvoir borner les erreurs d'échantillonnage (Marker 2001).

Nous notons que les règles de répartition susmentionnées sont faciles à appliquer dans tout contexte de sondage réel. Nous examinons les plans d'échantillonnage équilibrés qui respectent les tailles marginales d'échantillon définies par STDOM1 pour la première partition et par STDOM2 pour la deuxième : le plan BAL s'appuie sur les équations d'équilibrage (2.2.2) avec la spécification (2.3.2) du vecteur $\mathbf{z}_k$; les échantillons BALPOP satisfont (ou satisfont approximativement) les équations d'équilibrage $\sum_{k\in s}\pi_{k\ bd}\delta_k/\pi_k = n_{bd}\ \sum_{k\in s}\pi_{k\ bd}\delta_k/\pi_k = n_{bd}$ et $\sum_{k\in s\ bd}\delta_k/\pi_k = N_{bd}$ $(b = 1, ..., B$; $d = 1, ..., M_b)$. Nous avons obtenu les probabilités π_k des deux plans par la méthode simplifiée décrite à la section 3.3. En outre, pour achever la comparaison, nous avons considéré un plan coordonné (dénommé CPAR) obtenu en sélectionnant un seul échantillon pour chaque population marginale par échantillonnage de Pareto (Särndal et Lundström 2005) et en veillant à ce que le chevauchement des deux échantillons soit maximal. Les tailles marginales d'échantillon, définies respectivement par les plans STDOM1 et STDOM2, ne sont satisfaites qu'à titre d'espérance sur l'échantillonnage répété sous le plan CPAR ; les probabilités

d'inclusion sont calculées par la méthode itérative décrite dans Falorsi et coll. (2006). Nous avons sélectionné 500 échantillons Monte-Carlo pour chaque plan d'échantillonnage.

Pour chaque échantillon, nous avons calculé les estimations des totaux de domaine avec l'estimateur d'*Horvitz-Thompson* (HT), l'estimateur GREG *modifié* et l'estimateur *synthétique* exprimés sous la forme ${}_{bd}\hat{t}_{r,\text{syn}} = \sum_{k\in U_{bd}} \tilde{y}_{r,k}$. En ce qui concerne les estimateurs utilisant de l'information auxiliaire, nous avons exécuté deux modèles linéaires homoscédastiques simples, à savoir le modèle (6.1) qui utilise dix variables auxiliaires, dont six sont les indicateurs d'appartenance au *groupe d'activités économiques* et les quatre autres sont les indicateurs d'appartenance à la *classe de taille*, et le modèle (6.2), qui utilise les 44 variables indicatrices d'appartenance au domaine. Le modèle linéaire (6.1) s'exprime par

$$E_m(y_k) = \beta_h + \beta_j \text{ pour } k \in U_h \cap U_j, \qquad (6.1)$$

où U_h est la population d'entreprises du h^{e} (h = 1, ..., 6) *groupe d'activités économiques* et U_j est la population d'entreprises de la j^{e} (j = 1, ..., 4) *classe de taille* du nombre d'employés, et β_h et β_j sont les effets fixes du h^{e} groupe d'activités économiques de la j^{e} classe de taille.
Le modèle linéaire (6.2) est

$$E_m(y_k) = \beta_{1d} + \beta_{2d} \text{ pour } k \in U_{1d} \cap U_{2d}, \qquad (6.2)$$

où β_{1d} et β_{2d} sont les effets de domaine distincts.

Tableau 6.1
Plans d'échantillonnage utilisés dans l'étude par simulation

Plan d'échantillonnage	Abréviation
Stratifié par la partition 1 avec EASSR* dans chaque strate	STDOM1
Stratifié par la partition 2 avec EASSR* dans chaque strate	STDOM2
Échantillonnage équilibré sur les tailles marginales d'échantillon et sur les tailles de population	BALPOP
Échantillonnage équilibré sur les tailles marginales d'échantillon	BAL
Échantillonnage coordonné de Pareto	CPAR

*EASSR : échantillonnage aléatoire simple sans remise

Il convient de souligner que le but principal de l'expérience est de comparer différents plans d'échantillonnage en utilisant le même estimateur. Dans ce contexte, le choix du *meilleur modèle* n'est pas un problème essentiel ; par conséquent, nous avons considéré deux modèles possibles assez généraux qui peuvent être appliqués dans toutes les conditions des domaines planifiés. Le modèle (6.1) est un peu plus fiable que l'autre, parce que les estimations des paramètres de régression sont fondées sur de grandes tailles d'échantillon ; par contre, le modèle (6.2) permet d'évaluer l'effet de la planification des tailles d'échantillon de domaine, mais les estimations de chaque paramètre de régression sont fondée sur de petites tailles d'échantillon. Sous le modèle (6.2), les estimateurs synthétique et GREG modifié donnent des résultats identiques. Dans les équations qui suivent, chaque stratégie d'échantillonnage est indiquée en abrégé par le couple (plan, est.), où plan indique l'un des cinq plans d'échantillonnage mentionnés au tableau 6.1 et est. représente les estimateurs HT, synthétique et GREG modifié susmentionnés.

L'analyse qui suit est fondée sur l'ensemble de petits domaines. Nous avons calculé deux mesures de qualité, à savoir le *biais relatif absolu moyen* ($\overline{\text{BRA}}$) et l'*erreur quadratique moyenne relative* ($\overline{\text{EQMR}}$), exprimées par

$$\overline{\text{BRA}}_F(\text{plan, est.}) =$$

$$\frac{1}{\text{card}(F)} \sum_{bd\in F} \left| \frac{1}{500} \sum_{i=1}^{500} \left[{}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{plan}) - {}_{bd}t_r \right] \Big/ {}_{bd}t_r \right| \times 100,$$

$$\overline{\text{EQMR}}_F(\text{plan, est.}) =$$

$$\frac{1}{\text{card}(F)} \sum_{bd\in F} \left\{ \frac{1}{500} \sum_{i=1}^{500} \left[{}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{plan}) - {}_{bd}t_r \right]^2 \Big/ {}_{bd}t_r^2 \right\} \times 100$$

où F est un sous-ensemble particulier des domaines marginaux, card(F) est la cardinalité de F, ${}_{bd}\hat{t}^{i}_{r,\text{est}}(\text{plan})$ et la i^{e} estimation d'échantillon Monte-Carlo (i = 1, ..., 500) du total ${}_{bd}t_r$ en suivant la stratégie (plan, est.). En particulier, F représente alternativement le sous-ensemble de petits domaines de DOM1, celui de DOM2 ou l'ensemble global de petits domaines (de DOM1 ainsi que DOM2).

L'étude par simulation Monte-Carlo fait ressortir que les méthodes de stratification multidimensionnelle proposées dans le présent article permettent de contrôler le biais et la variabilité par rapport aux deux stratégies de référence (STDOM1 et STDOM2) en regroupant l'une des deux variables de stratification.

Les principaux résultats de l'expérience se rapportant à l'ensemble de petits domaines sont présentés au tableau 6.2. Ce dernier comprend quatre parties : la première illustre les mesures de qualité de l'estimateur HT, les deuxième et troisième sont consacrées, respectivement, aux estimateurs synthétique et GREG modifié basés sur dix variables auxiliaires (modèle (6.1)), tandis que la quatrième donne le résultat des estimateurs synthétique et GREG modifié basés sur les 44 variables indicatrices d'appartenance au domaine (modèle (6.2)). Nous limitons nos commentaires à la variable *valeur ajoutée*, mais des considérations semblables pourraient être exprimées pour la variable *coût de la main-d'œuvre*. En général, les commentaires se rapportent à l'ensemble global de petits domaines.

Tableau 6.2
Biais relatif absolu moyen ($\overline{BRA}$) et erreur quadratique moyenne relative ($\overline{EQMR}$) des stratégies d'échantillonnage de petit domaine

Plan d'échantillonnage	Valeur ajoutée						Coût de la main-d'œuvre					
	DOM1		DOM2		Global		DOM1		DOM2		Global	
	$\overline{BRA}$	$\overline{EQMR}$	$\overline{BRA}$	$\overline{EQMR}$	$\overline{BRA}$	$\overline{EQMR}$	$\overline{BRA}$	$\overline{EQMR}$	$\overline{BRA}$	$\overline{EQMR}$	$\overline{BRA}$	$\overline{EQMR}$
Estimateur d'Horvitz-Thompson (partie 1)												
STDOM1	1,79	43,19	8,18	148,28	5,41	102,74	1,72	42,82	6,86	155,87	4,63	106,88
STDOM2	3,42	107,49	0,47	15,26	1,75	55,23	3,32	105,66	0,46	12,66	1,70	52,96
BALPOP	0,77	24,86	1,29	38,49	1,06	32,58	0,74	23,60	1,20	34,26	1,00	29,64
BAL	0,84	25,43	1,45	40,61	1,19	34,03	0,79	24,22	1,57	35,80	1,23	30,78
CPAR	1,35	32,52	2,18	53,85	2,18	44,60	1,44	31,68	2,62	51,44	2,11	42,88
Estimateur synthétique avec 10 variables auxiliaires (partie 2)												
STDOM1	14,22	18,88	13,81	100,55	13,99	65,16	12,29	18,40	9,25	95,03	10,57	61,83
STDOM2	24,82	33,96	14,48	15,96	20,34	26,16	13,13	14,79	12,46	23,11	12,75	19,51
BALPOP	13,68	17,51	24,98	43,98	20,09	32,51	11,89	15,60	12,35	33,08	12,15	25,50
BAL	14,92	18,46	21,82	41,66	18,83	31,61	13,37	16,91	10,41	32,64	11,69	25,82
CPAR	13,68	17,83	23,45	44,63	19,22	33,02	11,82	16,13	11,69	34,93	11,75	26,78
Estimateur GREG modifié avec 10 variables auxiliaires (partie 3)												
STDOM1	2,35	30,13	11,26	119,95	7,40	81,03	1,86	29,28	11,79	119,23	7,49	80,25
STDOM2	3,98	58,62	0,95	15,26	2,26	34,05	2,90	52,66	0,93	12,66	1,78	29,99
BALPOP	1,11	19,41	2,20	25,80	1,73	23,03	1,01	16,42	1,99	21,73	1,57	19,43
BAL	1,63	19,41	1,76	26,11	1,70	23,21	1,21	16,72	2,08	21,96	1,70	19,69
CPAR	1,04	21,27	1,63	29,30	1,37	25,82	1,03	18,27	1,11	24,60	1,08	21,86
Estimateur synthétique ou GREG modifié avec 44 variables auxiliaires (partie 4)												
STDOM1	3,39	31,30	27,48	63,22	17,04	49,39	2,76	30,80	28,67	63,05	17,44	49,08
STDOM2	17,24	102,24	1,37	20,65	8,25	56,00	23,00	102,64	1,42	19,10	10,77	55,30
BALPOP	1,07	20,71	1,97	26,98	1,58	24,26	1,08	17,62	1,93	24,07	1,56	21,27
BAL	1,47	20,36	2,13	28,46	1,84	24,95	1,41	17,66	2,02	25,10	1,75	21,88
CPAR	1,79	23,38	2,22	32,39	2,03	28,48	1,65	20,73	2,08	30,39	1,90	26,21

En premier lieu, l'examen de l'estimateur HT nous mène aux constatations qui suivent.

- Les deux plans de référence (STDOM1 et STDOM2) ont une ($\overline{EQMR}$) pour les domaines non planifiés égale à 148,28 % et 107,49 %, respectivement. Ces valeurs sont la cause des grandes valeurs de l'($\overline{EQMR}$) calculées pour l'ensemble global de petits domaines, lesquelles sont égales à 102,74 % et 55,23 %, respectivement.
- Le plan STDOM2 donne de meilleurs résultats que le plan STDOM1. Cette constatation tient au fait que le critère de stratification de STDOM2 est corrélé aux variables d'intérêt et exerce un contrôle sur un plus grand nombre de petits domaines que la stratification STDOM1.
- En ce qui concerne l'ensemble global de petits domaines, le plan BALPOP est plus efficace, en ce qui a trait au $\overline{BRA}$ (1,06 %) et à l'$\overline{EQMR}$ (32,58 %), même si le plan BAL est à peine moins bon.
- La stratégie fondée sur l'échantillonnage coordonné produit des valeurs pires que l'échantillonnage équilibré, mais donne de meilleurs résultats en terme d'$\overline{EQMR}$ que les stratégies de référence.

Dans le cas de l'estimateur synthétique fondé sur dix variables auxiliaires, quelques questions se dégagent.

- Tous les plans sont caractérisés par un biais important. Le plan STDOM1 a un $\overline{BRA}$ égal à 13,99 % (bien que son $\overline{EQMR}$, qui est égale à 65,16 %, soit inacceptable). Les autres plans ont un $\overline{BRA}$ supérieur à 18 %. Ces résultats indiquent que l'estimateur synthétique ne devrait être utilisé qu'avec prudence.
- Le plan STDOM2 est celui dont l'$\overline{EQMR}$ est la plus faible (26,16 %), ce qui est dû à une forte réduction de la variance de DOM1. Cependant, son $\overline{BRA}$ (20,34 %) est plus grand que celui de tous les autres plans.
- En ce qui concerne le biais et la variance, le comportement des plans équilibrés et coordonné sont plus ou moins les mêmes. Le plan BAL produit les valeurs de $\overline{BRA}$ (18,33 %) et d'$\overline{EQMR}$ (31,61 %) les plus faibles.

Les résultats expérimentaux pour l'estimateur GREG modifié mènent aux commentaires qui suivent.

- Tous les plans révèlent une amélioration importante des mesures de qualité. En général, le $\overline{\text{BRA}}$ est remarquablement réduit comparativement au même indicateur calculé sur l'estimateur synthétique. Seul le plan STDOM1 produit encore un $\overline{\text{BRA}}$ élevé (7,40 %).
- Dans le cas du plan STDOM2, la réduction du biais est plus que compensée par l'accroissement de la variabilité, ce qui donne une $\overline{\text{EQMR}}$ égale à 34,05 %.
- Les plans équilibrés et coordonnés produisent tous de bons résultats, quoique les plans équilibrés soient un peu meilleurs, l'$\overline{\text{EQMR}}$ étant approximativement égale à 23 %.

Enfin, dans la quatrième partie, nous notons que les estimateurs synthétique ou GREG modifié basés sur les 44 variables auxiliaires produisent des résultats analogues à ceux de l'estimateur GREG modifié basé sur 10 variables auxiliaires. Les plans équilibrés sont les meilleurs, avec une légère préférence pour l'échantillonnage BALPOP.

En général, les plans équilibrés semblent représenter une bonne stratégie en vue de contrôler le biais et la variance de l'ensemble global de petits domaines.

La conclusion est que, dans les quatre parties du tableau, BALPOP est le plan qui donne généralement les meilleurs résultats globaux en ce qui concerne le biais et la précision. La stratégie basée sur le plan d'échantillonnage BALPOP conjugué à l'estimateur GREG modifié avec les dix variables auxiliaires (partie 3) est un choix prudent pour la valeur ajoutée ainsi que pour le coût de la main-d'œuvre. Le plan BAL donne aussi de bons résultats. En outre, le tableau indique que l'utilisation de l'estimateur synthétique de la partie 2 doit être envisagée avec prudence, car le biais peut être beaucoup plus important que prévu et que le carré du biais serait l'élément dominant de l'$\overline{\text{EQMR}}$

7. Conclusion

Le présent article illustre une stratégie d'échantillonnage efficace qui permet d'obtenir des tailles d'échantillon planifiées pour les domaines appartenant à différentes partitions de la population et de s'assurer que les erreurs d'échantillonnage des estimations de domaine soient inférieures à des seuils donnés. La stratégie d'échantillonnage, qui englobe le cas multivarié multidomaine, est utile quand la taille globale d'échantillon est bornée. Dans ces conditions, la solution standard, qui consiste à utiliser un échantillon stratifié dont les strates sont données par le recoupement des variables qui définissent les diverses partitions, n'est pas applicable, puisque le nombre de strates est plus grand que la taille globale d'échantillon.

La stratégie d'échantillonnage proposée est fondée sur l'utilisation d'une méthode d'échantillonnage équilibré et d'un estimateur de type GREG. Elle peut être étendue facilement à une stratégie comportant l'utilisation d'estimateurs de petit domaine directs et indirects.

L'un des principaux avantages de la solution proposée est son application facile, puisqu'elle est exécutée par des algorithmes qui sont fondés sur des outils logiciels gratuits ou qui peuvent être utilisés avec des tableurs automatisés. Toutefois, d'autres aspects intéressants se dégagent.

L'analyse empirique des données réelles sur les entreprises révèle de bonnes propriétés de la stratégie proposée, qui semble être robuste même lorsque l'on s'écarte des conditions idéales (autrement dit, les estimations semblent être de haute qualité, même si les probabilités d'inclusion de l'échantillon diffèrent des probabilités optimales). Ces résultats justifient des travaux supplémentaires afin d'établir de manière systématique les conditions sous lesquelles la méthode proposée donnera de bons résultats.

En outre, la stratégie proposée semble bien convenir pour les grands ensembles de données, en ce qui concerne le temps de calcul et, par conséquent, elle paraît appropriée pour les enquêtes à grande échelle.

Enfin, l'approche décrite représente une stratégie globale originale d'échantillonnage pour petits domaines qui tient compte simultanément du plan d'échantillonnage et de l'estimateur. L'article analyse en profondeur les questions liées au plan de sondage, mais d'autres travaux de recherche sont nécessaires pour étudier plus minutieusement les problèmes d'estimation. En particulier, les futurs travaux devraient être axés sur l'amélioration des estimateurs fondés sur un modèle ou assistés par un modèle due à la présence d'unités d'échantillonnage dans chaque domaine d'estimation, ce qui permettra d'utiliser des modèles contenant des effets de petit domaine spécifiques et produira des estimations plus précises des paramètres d'intérêt au niveau du petit domaine. Ces aspects semblent représenter un domaine d'étude prometteur.

Bibliographie

Bethel, J. (1989). Répartition de l'échantillon dans les enquêtes à plusieurs variables. *Techniques d'enquête*, 15, 49-60.

Bishop, Y., Fienberg, S. et Holland, P. (1975). *Discrete Multivariate Analysis*. MIT Press, Cambridge, MA.

Chauvet, G., et Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-62.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques.* New York : John Wiley & Sons, Inc.

Deville, J.-C., et Tillé, Y. (2000). Selection a several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86, 89-101.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., et Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Dykstra, R., et Wollan, P. (1987). Finding I-projections subject to a finite set of linear inequality constraints. *Applied Statistics*, 36, 377-383.

Falorsi, P.D., Orsini, D. et Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition*, 7, 1173-1198.

Falorsi, P.D., et Righi, P. (2008). An efficient multi-way design strategy for domain estimation. *Contributi ISTAT* (downloadable on http://www.istat.it/dati/pubbsci/contributi/).

Jessen, R.J. (1970). Probability sampling with marginal constraints. *Journal American Statistical Society*, 65, 776-795.

Lehtonen, R., et Veijanen, A. (1998). Estimateurs de régression généralisés logistiques. *Techniques d'enquête*, 24, 53-58.

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003). L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines. *Techniques d'enquête*, 1, 37-49.

Lu, W., et Sitter, R.R. (2002). Méthode pratique de stratification multiple par programmation linéaire. *Techniques d'enquête*, 2, 215-224.

Marker, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

Marker, D.A. (2001). Production d'estimations régionales d'après les données d'enquêtes nationals : méthodes visant à réduire au minimum l'emploi d'estimateurs indirects. *Techniques d'enquête*, 27, 201-207.

Montanari, G.E., et Ranalli, M.G. (2003). Nonparametric methods in survey sampling. Dans *New Developments in Classification and Data Analysis*, (Éds., M. Vinci, P. Monari, S. Mignani et A. Montanari), Springer, Berlin.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Särndal, C.-E., et Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York : Springer-Verlag.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Singh, M.P., Gambino, J. et Mantel, H.J. (1994). Les petites régions : problèmes et solutions. *Techniques d'enquête*, 20, 3-15.

Tillé, Y., et Favre, A.C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Valliant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference*: *A Prediction Approach*. New York : John Wiley & Sons, Inc.

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

Techniques d'enquête, décembre 2008
Vol. 34, N° 2, pp. 261-276
Statistique Canada, N° 12-001-X au catalogue

Estimation pour petits domaines sous un modèle à effets aléatoires en deux parties avec application à l'estimation de la littératie dans les pays en voie de développement

Danny Pfeffermann, Bénédicte Terryn et Fernando A.S. Moura [1]

Résumé

Le présent article traite de situations où la valeur de la réponse cible est zéro ou une observation issue d'une loi continue. Un exemple type analysé dans l'article est l'évaluation des compétences en littératie, dont le résultat peut être zéro, ce qui représente l'analphabétisme, ou un score positif mesurant le niveau de littératie. Nous cherchons à déterminer comment obtenir des estimations valides de la réponse moyenne ou de la proportion de réponses positives pour des petits domaines pour lesquels nous ne disposons que de petits échantillons, voire aucun. Comme dans d'autres problèmes d'estimation pour petits domaines, la petite taille de l'échantillon pour au moins certains domaines échantillonnés et/ou l'existence de domaines non échantillonnés obligent à utiliser des méthodes fondées sur un modèle. Toutefois, les méthodes disponibles ne conviennent pas pour ce genre de données, à cause de la distribution mixte des réponses caractérisée par un pic important à zéro juxtaposé à une distribution continue pour le reste des réponses. Par conséquent, nous élaborons un modèle à effets aléatoires en deux parties et montrons comment l'ajuster et évaluer la qualité de son ajustement, et comment calculer les estimations sur petits domaines d'intérêt et mesurer leur précision. Nous illustrons la méthode proposée au moyen de données simulées et de données provenant d'une enquête sur la littératie réalisée au Cambodge.

Mots clés : Intervalles de crédibilité ; modèle linéaire mixte généralisé ; qualité de l'ajustement ; modèle linéaire mixte ; MCMC ; biais de prédiction ; EQM de la prédiction.

1. Introduction

Nous examinons dans le présent article des situations où la valeur de la réponse cible est soit zéro, soit une observation issue d'une loi continue. Un exemple type analysé dans l'article est l'évaluation des compétences en littératie fondée sur un examen écrit dont le résultat peut être zéro, indiquant l'analphabétisme, ou un score positif compris dans une fourchette donnée mesurant le niveau de littératie. Un autre exemple est la consommation de drogues illicites (ou de certains aliments), où une valeur nulle signifie « aucune consommation », tandis qu'un résultat positif mesure la quantité consommée. Nous étudions le moyen d'obtenir des estimations valides de la réponse moyenne (niveau moyen de littératie dans notre exemple) ou de la proportion de réponses positives (proportion de personnes qui savent lire et écrire) dans des petits domaines pour lesquels nous ne disposons que de petits échantillons, voire aucun. Comme dans d'autres problèmes d'estimation pour petits domaines, la petite taille des échantillons dans les domaines échantillonnés et l'existence de domaines non échantillonnés obligent à appliquer des méthodes fondées sur un modèle.

Nous proposons l'utilisation d'un modèle à effets aléatoires en deux parties et montrons comment l'ajuster et évaluer la qualité de son ajustement, et comment obtenir les estimations sur petits domaines d'intérêt et mesurer leur précision. La première partie du modèle spécifie la probabilité qu'un score soit nul et la deuxième, la distribution des scores positifs. Le modèle n'est pas nouveau et est utilisé dans d'autres applications (voir, par exemple, Olsen et Schafer 2001, et la discussion et les références figurant dans cet article), mais autant que nous sachions, cette forme de distribution mixte n'a pas été étudiée auparavant dans la littérature traitant de l'estimation pour petits domaines. Notons que, dans notre application, les scores nuls sont des zéros « structurels » (vrais). Un ensemble connexe d'études traitent du nombre excessif de valeurs nulles dans les données de comptage, lequel peut résulter d'une combinaison de surdispersion et de surreprésentation réelle de zéros (true zero inflation). Les données présentant un excès de zéros (zero-inflated data) sont des données qui comportent une plus grande proportion de zéros que celle que l'on s'attendait à obtenir à partir de données de comptage (Poisson) pures. Voir, par exemple, Barry et Welsh (2002).

La première partie de notre modèle est la fonction logistique, utilisée pour modéliser la probabilité d'un score positif. La deuxième partie est un modèle linéaire contenant des termes d'erreur normaux ajustés aux réponses non nulles. Les deux modèles contiennent des covariables au niveau de l'individu et du domaine, ainsi que des effets

1. Danny Pfeffermann, Université hébreuse de Jérusalem, Israël et Southampton Statistical Sciences Research Institute, R.-U. Courriel : msdanny@huji.ac.il; Bénédicte Terryn, Department for International Development (DFID), R.-U.; Fernando A.S. Moura, Université fédérale de Rio de Janeiro, Brésil.

aléatoires de domaine reflétant les variations non expliquées par les covariables. Le modèle tient compte des corrélations entre les effets aléatoires correspondants des deux parties et est ajusté par application de simulations Monte-Carlo par chaîne de Markov (MCMC).

Nous avons ajusté le modèle en deux parties aux données recueillies dans le cadre de l'enquête nationale auprès des ménage sur la littératie réalisée au Cambodge en 1999, appelée « Évaluation des niveaux de littératie fonctionnels de la population adulte ». La figure 1 donne l'histogramme des scores de littératie observés pour cette enquête. Dans l'application décrite ici, nous produisons des estimations sur petits domaines pour les districts de résidence et les villages emboîtés, ce qui nécessite l'utilisation d'un modèle en deux parties à effets aléatoires à trois niveaux. Pour évaluer la qualité de l'ajustement du modèle, nous utilisons des statistiques descriptives simples et des données simulées à l'aide du modèle. L'utilisation de simulations nous permet aussi de comparer les résultats de l'ajustement du modèle en deux parties « complet » à ceux obtenus en ajustant les deux parties du modèle séparément, sans tenir compte des corrélations entre les effets aléatoires dans les deux parties. Nous les comparons également aux résultats obtenus en ignorant la nature particulière des données et en ajustant la partie linéaire de toutes les réponses, sans tenir compte de l'existence de nombreux scores nuls.

Dans la suite, afin de faciliter l'exposé et la discussion, nous examinons les scores de littératie mesurés pour les individus résidant dans les villages emboîtés dans les districts, mais comme nous l'avons mentionné plus haut, le modèle que nous étudions ici peut être utilisé pour de nombreuses autres applications.

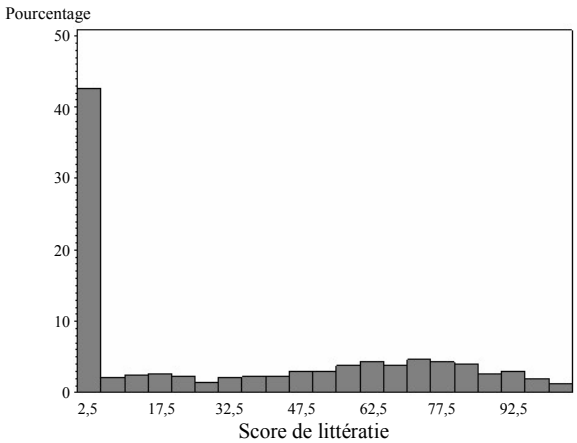


Figure 1 Histogramme des scores de littératie dans l'enquête nationale sur la littératie au Cambodge, 1999

2. Modèle et prédicteurs pour petits domaines

2.1 Le modèle en deux parties

Soit y la réponse (score du test de littératie dans notre application) et R les covariables et les effets aléatoires. Alors,

$$\begin{aligned} E(y|R=r) &= E(y|R=r,\ y=0)\ \Pr(y=0|R=r) \\ &\quad + E(y|R=r,\ y>0)\ \Pr(y>0|R=r) \\ &= E(y|R=r,\ y>0)\ \Pr(y>0|R=r). \quad (1) \end{aligned}$$

Dans le cas du problème d'estimation pour petits domaines étudié ici, nous appliquons un modèle emboîté à trois niveaux avec les districts de résidence définissant le premier niveau, les villages, le deuxième niveau et les individus, le troisième niveau. Pour l'individu k résidant dans le village j du district i, avec les covariables et les effets aléatoires $R_{ijk} = r$, nous avons par conséquent la relation

$$\begin{aligned} &E(y_{ijk}\,|\,R_{ijk}=r) = \\ &\qquad E(y_{ijk}\,|\,R_{ijk}=r,\ y_{ijk}>0)\ \Pr(y_{ijk}>0\,|\,R_{ijk}=r). \quad (2) \end{aligned}$$

Nous allons maintenant modéliser les deux parties du deuxième membre de (2). Pour les individus dont la réponse est *positive*, nous émettons l'hypothèse du « modèle linéaire mixte »,

$$\begin{aligned} &y_{ijk} = x'_{ijk}\beta + u_i + v_{ij} + \varepsilon_{ijk}; \\ &u_i \sim N(0,\ \sigma_u^2);\ v_{ij} \sim N(0,\ \sigma_v^2);\ \varepsilon_{ijk} \sim N(0,\ \sigma_\varepsilon^2), \quad (3) \end{aligned}$$

où x_{ijk} représente les valeurs des covariables au niveau de l'individu et du domaine, u_i est un effet de district aléatoire et v_{ij} est un effet de village aléatoire emboîté. Nous émettons l'hypothèse d'indépendance mutuelle des effets aléatoires u_i et v_{ij}, et des résidus ε_{ijk} entre les districts et villages, ainsi qu'à l'intérieur des districts et des villages. Ces effets et résidus tiennent compte de la variation des scores individuels non expliquée par les covariables, et définissent les corrélations entre les scores des individus résidant dans un même village et les corrélations entre les scores des individus résidant dans un même district, mais dans des villages différents.

$$\begin{aligned} &\mathrm{Corr}(y_{ijk},\ y_{i'j'k'}) = \\ &\begin{cases} (\sigma_u^2 + \sigma_v^2) / (\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2) & \text{si } i=i',\ j=j',\ k\neq k' \\ \sigma_u^2 / (\sigma_u^2 + \sigma_v^2 + \sigma_\varepsilon^2) & \text{si } i=i',\ j\neq j' \\ 0 & \text{si } i\neq i' \end{cases} . \quad (4) \end{aligned}$$

Pour les probabilités de réponses positives (la deuxième partie de l'équation (2)), nous émettons l'hypothèse du « modèle linéaire mixte généralisé »,

$$p_{ijk} = \Pr(y_{ijk} > 0 | x^*_{ijk}, u^*_i, v^*_{ij}) = \frac{\exp(x'^*_{ijk}\gamma + u^*_i + v^*_{ij})}{1 + \exp(x'^*_{ijk}\gamma + u^*_i + v^*_{ij})}$$

$$u^*_i \sim N(0, \sigma^2_{u^*}); \; v^*_{ij} \sim N(0, \sigma^2_{v^*}), \quad (5)$$

qui implique que $\text{logit}(p_{ijk}) = \log(p_{ijk}/(1-p_{ijk})) = x'^*_{ijk}\gamma + u^*_i + v^*_{ij}$. Ici, de nouveau, u^*_i et v^*_{ij} représentent des effets aléatoires indépendants de district et de village non expliqués par les covariables x^*_{ijk}. Notons que les covariables x_{ijk} dans l'équation (3) et les covariables x^*_{ijk} dans l'équation (5) peuvent différer (voir l'étude empirique à la section 4).

Remarque 1. On pourrait soutenir que le modèle linéaire mixte (3) avec les hypothèses de normalité supplémentaires implique un modèle probit correspondant pour les probabilités p_{ijk}. Toutefois, cela n'est pas le cas, car nous supposons uniquement que le modèle (3) s'applique aux scores positifs. Il s'ensuit que les paramètres des deux modèles peuvent être considérés comme distincts au sens de Rubin (1976).

Nous tenons compte des corrélations non nulles entre les effets aléatoires de district dans les deux parties, et faisons de même pour les effets aléatoires de village. Cette hypothèse est raisonnable, puisque, pour des valeurs données des covariables, on peut s'attendre à ce qu'un individu résidant dans un domaine caractérisé par des scores de littératie élevés ait une plus forte probabilité d'obtenir un score positif qu'un individu résidant dans un domaine où les scores sont faibles. La grandeur de ces corrélations et l'importance de leur prise en compte dans l'ajustement du modèle dépendent du pouvoir prédictif des covariables disponibles pour les deux parties du modèle, ou bien, des variances des effets aléatoires (les variances sont d'autant plus faibles que le pouvoir prédictif des covariables est élevé). Nous modélisons les corrélations en supposant que

$$u^*_i | u_i \sim N(K_u u_i, \sigma^2_{u^*|u}); \; v^*_{ij} | v_{ij} \sim N(K_v v_{ij}, \sigma^2_{v^*|v}). \quad (6)$$

La figure 2 fournit des preuves à l'appui de cette proposition obtenues à l'aide des données d'échantillon pour le centre du Cambodge qui sont utilisées dans l'étude empirique décrite à la section 4. (Les corrélations empiriques entre les variables mesurées sur les deux axes sont égales à 0,25 pour les villages et 0,38 pour les districts.)

2.2 Paramètres d'intérêt et prédicteurs

Pour le village (i, j) de taille N_{ij}, les paramètres d'intérêt du petit domaine sont la moyenne réelle des scores de littératie, $\bar{Y}_{ij} = \sum_{k=1}^{N_{ij}} y_{ijk} / N_{ij}$, et la proportion de scores positifs $P_{ij} = \sum_{k=1}^{N_{ij}} \text{I}(y_{ijk} > 0) / N_{ij}$, où $\text{I}(y_{ijk} > 0) = 1$ si $y_{ijk} > 0$, et 0 autrement. Notons que les moyennes sont calculées sur l'ensemble des individus, y compris ceux dont le score est nul.

Sous le modèle (2), la moyenne est prédite par

$$\hat{\bar{Y}}_{ij} = \left[\sum_{k \in S_{ij}} y_{ijk} + \sum_{k \notin S_{ij}} \hat{y}_{ijk} \right] \Big/ N_{ij}, \quad (7)$$

où S_{ij} définit l'échantillon provenant du village (i, j). En vertu de (3) et (5), les scores manquants peuvent être prédits sous l'approche fréquentiste par

$$\hat{y}_{ijk} = \frac{\exp(x'^*_{ijk}\hat{\gamma} + \hat{u}^*_i + \hat{v}^*_{ij})}{1 + \exp(x'^*_{ijk}\hat{\gamma} + \hat{u}^*_i + \hat{v}^*_{ij})} \times [x'_{ijk}\hat{\beta} + \hat{u}_i + \hat{v}_{ij}], \quad (8)$$

où $\hat{\beta}$, $\hat{\gamma}$, $\hat{u}_i$, $\hat{v}_i$, $\hat{u}^*_i$, $\hat{v}^*_{ij}$ désignent les estimations d'échantillon appropriées (voir la section suivante). Nous pourrions ajouter à la moyenne estimée, $(x'_{ijk}\hat{\beta} + \hat{u}_i + \hat{v}_{ij})$, une estimation $\hat{\varepsilon}_{ijk}$ obtenue par tirage à partir de la loi $N(0, \hat{\sigma}^2_\varepsilon)$ ou par sélection aléatoire d'un résidu estimé, $\hat{\varepsilon}_{i', j', k'} = (y_{i', j', k'} - x'_{i'j'k'}\hat{\beta} - \hat{u}_{i'} - \hat{v}_{i'j'})$, parmi les résidus estimés pour les individus échantillonnés. L'ajout d'estimations $\hat{\varepsilon}_{ijk}$ aux valeurs estimées de la moyenne reflète plus fidèlement la variabilité des réponses positives. Sous l'approche bayésienne, les scores manquants sont prédits par tirage aléatoire à partir de leur loi prédictive (voir la section suivante).

En vertu de (5), la proportion P_{ij} est prédite sous l'approche fréquentiste par

$$\hat{P}_{ij} = \frac{1}{N_{ij}} \left[\sum_{k \in S_{ij}} \text{I}(y_{ijk} > 0) + \sum_{k \notin S_{ij}} \hat{\text{I}}(y_{ijk} > 0) \right], \quad (9)$$

où

$$\hat{\text{I}}(y_{ijk} > 0) = \frac{\exp(x'^*_{ijk}\hat{\gamma} + \hat{u}^*_i + \hat{v}^*_{ij})}{1 + \exp(x'^*_{ijk}\hat{\gamma} + \hat{u}^*_i + \hat{v}^*_{ij})}.$$

Une solution bayésienne consiste à prédire les indicateurs $\text{I}(y_{ijk} > 0)$ par tirage aléatoire à partir de leur loi prédictive.

Les moyennes et les proportions de district sont prédites de manière analogue, ce qui revient à calculer la moyenne pondérée des prédicteurs correspondants de village, avec les poids définis par les tailles relatives de village.

Remarque 2. Le calcul du prédicteur défini par (7) et (8) requiert la connaissance des covariables x_{ijk}, x^*_{ijk} pour chaque unité de la population. De même, le calcul du prédicteur (9) requiert la connaissance des covariables x^*_{ijk} pour chaque unité de la population. Il en est généralement ainsi pour tous les modèles linéaires mixtes généralisés. L'information sur les covariables auxiliaires provient souvent de recensements ou d'autres dossiers administratifs. En l'absence de cette information, les covariables manquantes peuvent être imputées par tirage aléatoire à partir de leur loi paramétrique estimée ou de leur loi empirique.

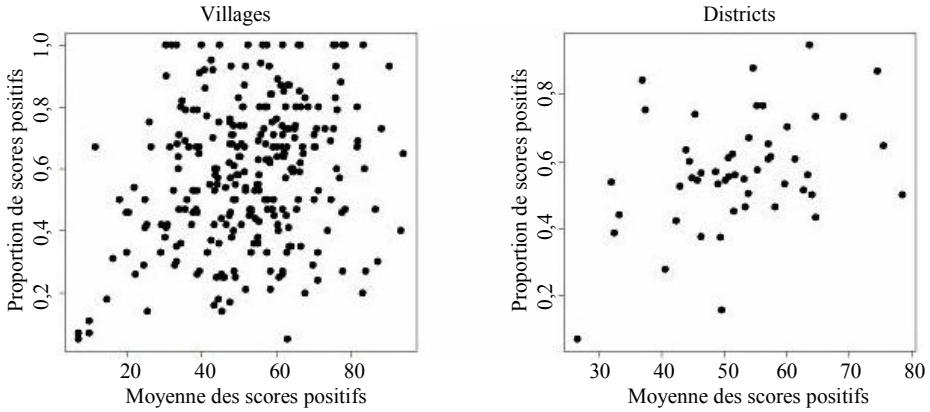


Figure 2 Proportion de scores positifs en fonction de la moyenne des scores positifs pour les districts et pour les villages du centre du Cambodge. Enquête nationale sur la littératie, 1999

3. Inférence

L'utilisation des prédicteurs sur petits domaines définis par les équations (7) à (9) requiert l'estimation des paramètres fixes (hyperparamètres) (β, σ_u^2, σ_v^2, σ_ε^2) de la partie linéaire (équation 3) et des paramètres fixes (γ, K_u, K_v, $\sigma^2_{u^*|u}$, $\sigma^2_{v^*|v}$) de la partie logistique (équations 5 et 6), ainsi que la prédiction des effets aléatoires $\{(u_i, v_{ij}; u_i^*, v_{ij}^*)\}$. Au cours des deux dernières décennies, des méthodes d'estimation des effets fixes et aléatoires quand on ajuste des modèles linéaires mixtes, ou des modèles linéaires mixtes généralisés seulement, ont été élaborées sous le paradigme tant fréquentiste que bayésien. Ces méthodes permettent aussi de calculer des estimateurs de l'erreur quadratique moyenne (EQM) ou du risque bayésien des prédicteurs sur petits domaines qui tiennent compte de l'estimation des hyperparamètres pour corriger l'ordre. Voir le livre publié par Rao (2003) et l'article plus récent de Jiang et Lahiri (2005) pour une revue et une discussion approfondies. Néanmoins, l'étude du modèle en deux parties défini par les équations (2) à (6) n'a pas été décrite dans la littérature sur les petits domaines et nous allons maintenant examiner quelques possibilités d'ajustement de ce modèle.

3.1 Inférence fondée sur la vraisemblance complète

Définissons $I_{ijk} = 1(0)$ si $Y_{ijk} > 0 (= 0)$ et désignons $r_{ijk} = (x_{ijk}, u_i, v_{ij})$, $r^*_{ijk} = (x^*_{ijk}, u^*_i, v^*_{ij})$. Pour des vecteurs donnés r_{ijk}, r^*_{ijk}, la vraisemblance du modèle en deux parties prend la forme

$$L = \prod_{i,\, j,\, k \in s} (p_{ijk})^{I_{ijk}} [f(y_{ijk} | r_{ijk},\ y_{ijk} > 0)]^{I_{ijk}} (1 - p_{ijk})^{(1 - I_{ijk})}, \quad (10)$$

où $s = \cup s_{ij}$ désigne l'échantillon provenant de tous les villages, p_{ijk} est définie par (5) et $f(y_{ijk} | r_{ijk}, y_{ijk} > 0)$ est la densité normale de moyenne $(x'_{ijk}\beta + u_i + v_{ij})$ et de variance σ_ε^2 (équation 3). L'utilisation de cette vraisemblance pour l'inférence pose toutefois des problèmes, parce qu'en fait, les effets aléatoires $\{(u_i, v_{ij}\ ;\ u_i^*, v_{ij}^*)\}$ sont inobservables. Par conséquent, une possibilité consiste à intégrer la vraisemblance sur la distribution (normale) conjointe des effets aléatoires, telle qu'elle est définie par (3), (5) et (6), et à maximiser la vraisemblance intégrée par rapport aux (hyper) paramètres fixes (β, σ_u^2, σ_v^2, σ_ε^2) et (γ, K_u, K_v, $\sigma^2_{u^*|u}$, $\sigma^2_{v^*|v}$). Ayant estimé les paramètres fixes, nous pouvons prédire les effets aléatoires par leurs valeurs prévues sachant les données (en maintenant les estimations du maximum de vraisemblance fixes), ce qui requiert un autre ensemble d'intégrations. Olsen et Schafer (2001) considèrent un modèle en deux parties pour l'ajustement de données longitudinales et substituent à la vraisemblance intégrée une approximation de Laplace multivariée d'ordre élevé (Raudenbush, Yang et Yosef 2000). Pour calculer les prédicteurs bayésiens empiriques des effets aléatoires, ils utilisent l'échantillonnage d'importance (Tanner 1996), en donnant pour valeur aux paramètres fixes leurs estimations du maximum de vraisemblance. Toutefois, l'application de cette procédure comprend des calculs très compliqués et les estimateurs de l'erreur quadratique moyenne (EQM) des prédicteurs sur petits domaines obtenus de cette façon ne tiennent pas compte de la variation induite par l'estimation des paramètres fixes. En général, la contribution de l'estimation des paramètres fixes à l'EQM totale ne peut pas être ignorée, à moins que les nombres de districts et de villages échantillonnés soient très grands.

3.2 Ajustement séparé des parties du modèle

L'idée ici est d'ajuster les deux parties du modèle séparément, puis de combiner les estimations pour calculer un prédicteur sur petits domaines défini par (7) et (8). Le prédicteur donné par (9) s'obtient directement de l'ajustement de la deuxième partie uniquement. Comme nous l'avons mentionné plus haut, les études traitant de l'ajustement des

parties séparées du modèle sont nombreuses et des logiciels sont disponibles, particulièrement pour les modèles linéaires mixtes. Il importe de souligner à cet égard que, sous le présent modèle en deux parties, les prédicteurs (7) à (9) sont des fonctions non linéaires des données et que, même si les hyperparamètres sont connus, il n'existe aucune formule explicite pour l'EQM des prédictions. Estimer l'EQM sous l'approche fréquentiste avec un biais d'ordre faible requiert par conséquent l'élaboration de nouvelles approximations appropriées ou des méthodes de rééchantillonnage, qui, dans le cas du prédicteur $\hat{\bar{Y}}_{ij}$ défini par (7) et (8), tiennent compte des corrélations entre les données dans les deux parties. Vient encore compliquer la situation le fait qu'en ajustant les deux parties séparément, la façon d'estimer les coefficients (K_u, K_v) qui définissent les corrélations entre les effets aléatoires dans les deux parties (équation 6) n'est pas claire. Une méthode jackknife d'estimation de l'EQM des prédictions pour le prédicteur donné par (9) sous ajustement de modèles séparés a été élaborée par Jiang, Lahiri et Wan (2002). Les estimateurs bootstrap applicables à ce prédicteur, de nouveau sous ajustement de modèles séparés, sont étudiés dans Hall et Maiti (2006).

3.3 Inférence bayésienne sous le modèle en deux parties

L'utilisation de méthodes bayésiennes requiert la spécification des lois a priori pour les paramètres fixes qui sous-tendent le modèle en deux parties (les coefficients β, γ, K_u, K_v et les variances σ_u^2, σ_v^2, σ_ε^2, $\sigma_{u|u}^2$, $\sigma_{v|v}^2$). Cependant, avec l'aide de simulations Monte-Carlo par chaîne de Markov (MCMC), cette approche permet l'échantillonnage à partir de la loi a posteriori des paramètres fixes et des effets aléatoires, donc l'échantillonnage à partir de la loi prédictive des réponses inobservées. Par conséquent, l'utilisation de cette approche donne la loi a posteriori complète des paramètres de petit domaine d'intérêt, ce qui permet le calcul de mesures correctes de l'EQM (variance a posteriori) ou d'intervalles de confiance (crédibilité) qui tiennent compte de toutes les sources de variation. Comme nous l'avons exposé plus haut, l'estimation de l'EQM des prédictions sous les approches antérieures pose des problèmes, particulièrement en ce qui concerne le prédicteur $\hat{\bar{Y}}_{ij}$ défini par (7) et (8). Un logiciel existe pour exécuter tous les calculs nécessaires, mais dans le cas de modèles complexes, ces calculs peuvent être intensifs et prendre beaucoup de temps.

Dans l'étude empirique décrite dans le présent article, nous avons suivi une approche bayésienne en utilisant le logiciel WinBUGS (Spiegelhalter, Thomas et Best 2003). Ce logiciel a la réputation d'être « convivial » et, si nous nous en tenons à notre expérience passée, il fonctionne très bien. Il existe évidemment de nombreux autres logiciels pour l'exécution de simulations MCMC, tels que MLwiN (Rasbash, Browne, Goldstein, Yang, Plewis, Healy, Woodhouse, Draper, Langford et Lewis 2002) ou R (Development Core Team 2008).

WinBUGS implémente l'algorithme MCMC avec l'échantillonneur de Gibbs (Gelfand et Smith 1990). Ce dernier échantillonne alternativement à partir de la loi conditionnelle de chacun des paramètres fixes et aléatoires (effets aléatoires), sachant les données et les autres paramètres. Il définit une chaîne markovienne, qui, sous certaines conditions de régularité, converge vers une réalisation de la loi a posteriori conjointe de tous les paramètres du modèle. Donc, à la fin du processus d'échantillonnage (à la convergence), l'algorithme produit une réalisation (unique) de chacun des paramètres fixes et aléatoires à partir de leur loi a posteriori conjointe, sachant les données. Les réalisations sont représentées ci-après par un tilde au-dessus des symboles. Les réalisations $\tilde{y}_{ijk}$ provenant de la loi a posteriori de y_{ijk} sont obtenues par tirage aléatoire de $I = 1$ (ou 0) avec les probabilités $\tilde{p}_{ijk}$ (ou $1 - \tilde{p}_{ijk}$) ; $\tilde{p}_{ijk} = \exp(x_{ijk}'^{*}\tilde{\gamma} + \tilde{u}_i^{*} + \tilde{v}_{ij}^{*}) \times [1 + \exp(x_{ijk}'^{*}\tilde{\gamma} + \tilde{u}_i^{*} + \tilde{v}_{ij}^{*})]^{-1}$, et en définissant

$$\tilde{y}_{ijk} = (x_{ijk}'\tilde{\beta} + \tilde{u}_i + \tilde{v}_{ij} + \tilde{\varepsilon}_{ijk}) \times \tilde{I}_{ijk}. \qquad (11)$$

La substitution de $\tilde{y}_{ijk}$ à $\hat{y}_{ijk}$ dans (7) et de $\tilde{I}_{ijk}$ à $\hat{I}_{ijk}$ dans (9) produit une valeur échantillonnée unique de la moyenne $\bar{Y}_{ij}$ et de la proportion P_{ij} à partir de leurs lois a posteriori respectives, pour chaque village (i, j). La répétition indépendante du même processus un grand nombre de fois (en utilisant des chaînes parallèles, voir plus bas) donne une approximation empirique de la loi a posteriori de la moyenne et de la proportion. Les moyennes réelles de village sont alors prédites en calculant la moyenne des valeurs échantillonnées correspondantes dans toutes les chaînes et les proportions de village sont obtenues de la même façon. L'EQM (risque bayésien) est estimée en calculant la variance empirique des valeurs échantillonnées. Les intervalles de crédibilité (confiance) avec le taux de couverture de $(1 - \alpha)$ sont définis par les quantiles de niveau $\alpha/2$ et $(1 - \alpha/2)$ de la loi a posteriori empirique. La même procédure est appliquée pour prédire les moyennes et les proportions de district, et pour calculer les variances et les intervalles de crédibilité de prédiction correspondants.

En pratique, l'utilisation de chaînes parallèles pour produire des réalisations indépendantes à partir de lois a posteriori prend généralement trop de temps, auquel cas l'échantillon peut être généré à partir d'une seule longue chaîne ou de quelques chaînes, mais en sélectionnant seulement toutes les r^e valeurs échantillonnées (après la convergence), ce qui réduit autant que possible les corrélations entre les valeurs échantillonnées adjacentes.

4. Résultats empiriques

4.1 Données et modèle

Pour les illustrations empiriques, nous utilisons des données provenant de l'enquête de 1999 « Évaluation des niveaux de littératie fonctionnels de la population adulte » au Cambodge. Il s'agit d'une enquête auprès des ménages dans le cadre de laquelle 6 548 adultes ont été interviewés et ont subi un test de littératie comprenant 20 tâches en langue Khmer pour lequel les scores variaient de 0 à 100 (voir la figure 1 dans l'introduction). L'enquête a été réalisée selon un plan d'échantillonnage stratifié à plusieurs degrés dont les strates étaient définies par les 24 provinces formant le pays. Chaque province est divisée en districts et environ la moitié de ceux-ci ont été sélectionnés dans l'échantillon (soit, en tout, 96 des 184 districts du pays). Deux communes ont été échantillonnées dans chacun des districts sélectionnés et, à part dans quelques cas, trois villages ont été sélectionnés dans chacune des communes échantillonnées. Enfin, des ménages ont été échantillonnés dans chaque village et un adulte a été sélectionné dans chaque ménage, en alternant en fonction de l'âge et du sexe. À chaque degré, la sélection a été faite selon un plan d'échantillonnage systématique. Le nombre de ménages sélectionnés par village était le même pour tous les villages appartenant à une même province. Les tailles totales des échantillons provinciaux ont été réparties proportionnellement aux tailles des populations provinciales.

Les petits domaines d'intérêt sont les districts et les villages. Nous limitons la présente étude aux 50 districts ruraux échantillonnés dans les provinces du centre du pays, pour lesquels le même modèle devrait être vérifié. De ces 50 districts, 5 avaient un échantillon de 20 adultes ou moins, et les 45 autres, un échantillon de 41 à 120 adultes. Dans l'ensemble de données réduit, le nombre de villages était 286, avec 47 villages pour lesquels la taille de l'échantillon était de 9 adultes ou moins et 193 villages pour lesquels la taille de l'échantillon était de 10 à 19 adultes. Le nombre total d'adultes dans l'échantillon est $n = 4\,028$.

Le tableau 1 donne les résultats obtenus en ajustant le modèle en deux parties complet aux données d'échantillon, en utilisant la méthode bayésienne et le logiciel décrit à la section 3.3. Dans les deux modèles, les covariables (variables explicatives) ont été sélectionnées par application de certaines procédures standard de sélection d'un modèle. Toutes les covariables, excepté l'âge, le niveau d'études et la taille du ménage, sont des variables indicatrices, qui prennent la valeur 1 quand la définition de la variable est satisfaite. Nous avons utilisé des lois a priori normales avec des grandes variances pour les éléments des vecteurs de coefficients β, γ et des lois a priori uniformes avec intervalle grand (mais fini) pour les écarts-types qui sous-tendent les deux parties du modèle et les coefficients K_u et K_v dans l'équation 6. Par défaut, WinBUGS sélectionne automatiquement la méthode d'échantillonnage de la loi conditionnelle de chacun des paramètres fixes et aléatoires en appliquant l'échantillonneur de Gibbs. Notons que les lois conditionnelles n'ont pas de forme explicite sous le présent modèle complet. Le logiciel sélectionne une méthode d'acceptation/rejet pour la partie logistique et d'échantillonnage par tranches (*slice sampling*) (Neal 2000) pour la plupart des autres paramètres et effet aléatoires.

Pour les simulations MCMC, nous avons produit une chaîne de longueur 50 000, avons écarté les 5 000 premières valeurs échantillonnées à titre d'« apprentissage », puis élagué la chaîne en prenant toutes les 150^{e} valeurs d'échantillon. Nous avons constaté, au moyen de certaines techniques graphiques d'usage répandu, que le rejet des 5 000 premières valeurs échantillonnées suffisait à garantir la convergence de la chaîne. Ces techniques comprennent la comparaison des histogrammes des lois a posteriori des divers paramètres basés sur différentes sous-séquences de la chaîne, l'inspection des traces de plusieurs chaînes simulées en parallèle, chacune avec une valeur de départ différente pour confirmer la stabilisation de la chaîne, et la représentation graphique des autocorrélations des valeurs échantillonnées pour confirmer l'indépendance après un sous-échantillonnage approprié. Voir Gamerman et Lopes (2006) pour une discussion plus approfondie et des exemples, y compris des tests de convergence plus formels. Soulignons aussi que les résultats des simulations de la section 4, obtenus en utilisant le modèle ajusté aux données réelles et en produisant une chaîne distincte de longueur 50 000 pour chaque simulation et en écartant les 5 000 premières valeurs à titre d'« apprentissage », sont très satisfaisants, ce qui constitue un autre indice de la convergence de la chaîne après les 5 000 premières valeurs.

Les coefficients K et les variances estimées des effets aléatoires impliquent $\widehat{\text{Corr}}(u_i, u_i^*) = 0{,}45$; $\widehat{\text{Corr}}(v_{ij}, v_{ij}^*) = 0{,}21$. Fait intéressant, les corrélations sont proches des corrélations empiriques obtenues en utilisant les moyennes brutes présentées à la fin de la section 2.1.

Les principaux résultats qui se dégagent du tableau 1 se résument comme il suit. Les coefficients de toutes les variables explicatives sont fortement significatifs (sur la base d'un test t standard) et possèdent généralement le signe attendu. D'autres variables que nous avions envisagé d'inclure dans les deux modèles se sont avérées non significatives. Les variances des effets aléatoires sont hautement significatives dans les deux modèles, ce qui indique leur contribution à l'explication de la variation des scores ou des probabilités de score positif qui n'est pas expliquée par les covariables incluses dans les deux modèles.

À titre de diagnostic supplémentaire pour le modèle logistique mixte, nous montrons à la figure 3 un diagramme de dispersion des proportions de scores positifs observées

$(I_{ijk} = 1)$ en fonction de la moyenne des probabilités prédites de score positif sous le modèle, dans le groupe de 50 individus définis par les valeurs ordonnées des probabilités prédites. Les valeurs tracées se situent presque sur une droite, ce qui témoigne d'un bon ajustement. La figure 4 donne un histogramme des résidus normalisés estimés de la partie linéaire mixte, $\hat{z}_{ijk} = \hat{\varepsilon}_{ijk} / \mathrm{ET}(\hat{\varepsilon}_{ijk}) = (y_{ijk} - x'_{ijk}\hat{\beta} - \hat{u}_i - \hat{v}_{ij}) / \mathrm{ET}(\hat{\varepsilon}_{ijk})$, où $\mathrm{ET}(\hat{\varepsilon}_{ijk})$ est l'écart-type empirique des résidus estimés. Bien qu'il ne s'agisse pas d'une courbe en cloche « parfaite », l'histogramme ne révèle aucun écart important par rapport à la loi normale.

Tableau 1
Paramètres estimés et erreurs types (e.t.) pour l'ajustement du modèle en deux parties

	Partie linéaire		Partie logistique	
Variables explicatives	Estimation	E.T.	Estimation	E.T.
Constante	$\hat{\beta}_0 = 6{,}90$	4,00	$\hat{\gamma}_0 = -6{,}48$	0,58
Années d'études	$\hat{\beta}_1 = 7{,}28$	0,53	$\hat{\gamma}_1 = 2{,}16$	0,12
Années d'études^2	$\hat{\beta}_2 = -0{,}24$	0,05	$\hat{\gamma}_2 = -0{,}13$	0,01
A participé à un programme d'alphabétisation	-	-	$\hat{\gamma}_3 = 2{,}44$	0,27
Aidé par l'intervieweur	-	-	$\hat{\gamma}_4 = 2{,}00$	0,17
Faible revenu	$\hat{\beta}_5 = -2{,}61$	0,88	$\hat{\gamma}_5 = -0{,}35$	0,14
Fonctionnaire/professionnel	$\hat{\beta}_6 = 13{,}91$	1,89	-	-
Sexe (1 pour féminin)	$\hat{\beta}_7 = -1{,}60$	0,81	$\hat{\gamma}_7 = -0{,}59$	0,14
Taille du ménage (adultes)	$\hat{\beta}_8 = 0{,}94$	0,29	-	-
Âge	$\hat{\beta}_9 = 0{,}84$	0,16	$\hat{\gamma}_9 = 0{,}14$	0,02
Âge^2	$\hat{\beta}_{10} = -0{,}01$	0,002	$\hat{\gamma}_{10} = -0{,}002$	0,00
Variances	Estimation	E.T.	Estimation	E.T.
Entre districts	$\hat{\sigma}_u^2 = 66{,}31$	16,72	$\hat{\sigma}_{u^*}^2 = 1{,}28$	0,34
Entre villages	$\hat{\sigma}_v^2 = 66{,}58$	10,45	$\hat{\sigma}_{v^*}^2 = 0{,}86$	0,19
Résidus	$\hat{\sigma}_\varepsilon^2 = 322{,}0$	10,12	-	-
Coefficients K (équation 6)*	Estimation		E.T.	
Effets aléatoires de district	$\hat{K}_u = 0{,}06$		0,02	
Effets aléatoires de village	$\hat{K}_v = 0{,}02$		0,01	

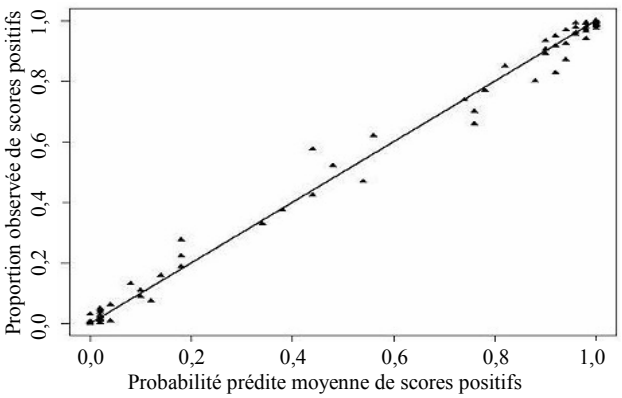


Figure 3 Probabilités observées et prédites de scores positifs

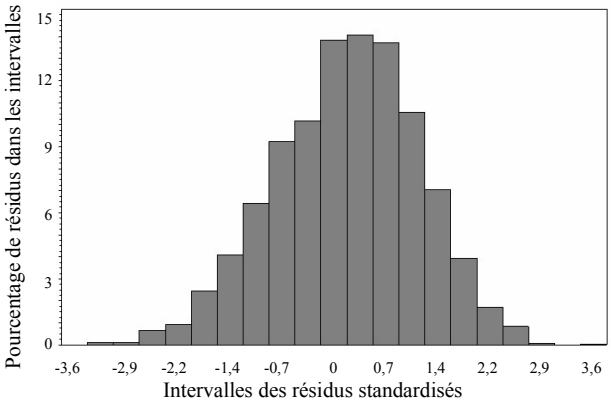


Figure 4 Histogramme des résidus standardisés pour la partie linéaire

En guise d'évaluation finale de la qualité de l'ajustement du modèle en deux parties, nous avons produit 200 nouveaux ensembles de données de taille $n = 4\,028$ au moyen du modèle en deux parties estimé du tableau 1, en utilisant les mêmes covariables que pour l'échantillon original. Les scores de test ont été produits en générant les effets aléatoires et les résidus $(u_i, v_{ij}, \varepsilon_{ijk})$ avec les variances estimées $(\hat{\sigma}^2_u, \hat{\sigma}^2_v, \hat{\sigma}^2_\varepsilon)$ (équation 3), en générant les effets aléatoires (u^*_i, v^*_{ij}) en utilisant l'équation (6) avec les coefficients et les variances estimés $(\hat{K}_u, \hat{\sigma}^2_{u^*|u}, \hat{K}_v, \hat{\sigma}^2_{v^*|v})$ en tirant aléatoirement des valeurs 1 ou 0 avec les probabilités $\Pr(I_{ijk}=1) = \exp(x'^*_{ijk}\gamma + u^*_i + v^*_{ij}) \times [1 + \exp(x'^*_{ijk}\gamma + u^*_i + v^*_{ij})]^{-1}$ (équation 5) et, dans le cas des valeurs 1, en produisant des scores non nuls $y_{ijk} = x'_{ijk}\hat{\beta} + u_i + v_{ij} + \varepsilon_{ijk}$ (équation 3). Nous avons calculé la variance $\hat{\sigma}^2_{u^*|u}$ sous la forme $\hat{\sigma}^2_{u^*|u} = (\hat{\sigma}^2_{u^*} - \hat{K}^2_u \hat{\sigma}^2_u)$, et avons calculé $\hat{\sigma}^2_{v^*|v}$ de la même façon (équation 6). Puis, pour chaque ensemble de données, nous avons calculé les moyennes et les proportions des scores pour chaque village et district, et nous les avons utilisées pour calculer les intervalles de confiance empiriques basés sur les 200 moyennes et proportions. Le tableau 2 donne les proportions de fois que les intervalles de confiance (IC) empiriques contiennent les valeurs correspondantes observées dans l'échantillon de l'enquête cambodgienne.

Les résultats du tableau 2 révèlent des taux de couverture très proches des valeurs nominales pour les villages, mais une sous-couverture allant jusqu'à 10 % pour les districts, ce qui tient probablement au fait que les derniers taux ne sont basés que sur 50 districts seulement.

4.2 Étude par simulation

L'expérience de simulation a pour but d'étudier l'efficacité du modèle en deux parties afin de produire des prédicteurs sur petits domaines et les mesures connexes des erreurs de prédiction. Elle permet également de comparer les résultats obtenus sous ce modèle à ceux obtenus lorsque les deux parties du modèle sont ajustées séparément, en ne tenant pas compte des corrélations entre les effets aléatoires correspondants dans les deux parties, ainsi qu'aux résultats obtenus lorsqu'on ajuste un modèle linéaire mixte à toutes les réponses, en ne tenant pas compte de l'accumulation de scores nuls. À cette fin, nous avons produit 300 nouvelles populations de $N = 4\,028$ scores et 300 nouveaux échantillons de taille $n = 1\,026$, de la même façon que les ensembles de données utilisés pour recalculer les intervalles de confiance présentés au tableau 2, mais au moyen d'un modèle contenant un moins grand nombre de variables explicatives que celui présenté au tableau 1. Dans la partie logistique, nous avons inclus quatre variables explicatives, à savoir le « nombre d'années d'études », la « participation à un programme d'alphabétisation », l'« obtention d'aide de l'intervieweur » et le fait d'avoir un « faible revenu ». Dans la partie linéaire, nous avons inclus cinq variables explicatives, à savoir le « nombre d'années d'études, le « sexe », la « taille du ménage », l'« âge » et l'« âge^2 ». Afin de fixer les valeurs des paramètres, nous avons ajusté séparément la partie linéaire et la partie logistique contenant les plus petits nombres de variables explicatives aux données d'échantillon originales. Nous avons fixé les corrélations entre les effets aléatoires des parties logistique et linéaire à 0,5 au niveau du district ainsi que du village.

Nous avons considéré les moyennes (proportions) de district et de village dans les populations simulées comme étant les moyennes (proportions) réelles de district et de village, ce qui nous permet d'évaluer la performance de divers prédicteurs. Comme nous l'avons souligné dans la remarque 2 à la section 2.2, la prédiction des moyennes et des proportions de petit domaine sous le modèle en deux parties requiert la connaissance des covariables pour toutes les unités de la population. Cette exigence est satisfaite dans l'étude par simulation, puisque les variables explicatives de l'échantillon original de 4 028 individus sont utilisées dans les populations simulées. Afin de spécifier les valeurs échantillonnées pour les variables explicatives, nous avons échantillonné 1 026 individus et utilisé les variables explicatives échantillonnées pour chacun des 300 échantillons. La moitié des individus dans la moitié des 286 villages ont été inclus dans l'échantillon, à l'exception des villages pour lesquels l'ensemble de données original comptait moins de cinq adultes, auquel cas tous les individus ont été échantillonnés. Nous avons appliqué ce critère de taille minimale afin d'éviter des problèmes de calcul durant l'exécution des simulations (voir la section 4.3). L'échantillon contenait des individus provenant de chacun des 50 districts ruraux, parmi lesquels 1 district avait une taille d'échantillon de 4, 4 districts avaient une taille d'échantillon de 9, 17 districts avaient une taille d'échantillon de $15 \le n_d \le 20$, et les 28 autres avaient une taille d'échantillon de $21 \le n_d \le 30$. Comme nous l'avons mentionné plus haut, l'échantillon contenait les individus provenant de la moitié des 286 villages, dans laquelle 29 villages avaient une taille d'échantillon $2 \le n_v \le 5$, 109 avaient une taille d'échantillon $6 \le n_v \le 10$, et 5 avaient une taille d'échantillon $11 \le n_v \le 19$.

Tableau 2
Proportions de fois que les intervalles de confiance empiriques contiennent les moyennes et les proportions observées dans l'échantillon

	IC à 90 % empirique		IC à 95 % empirique	
Petites domaines	Districts	Villages	Districts	Villages
Nombre de petits domaines	50	286	50	286
% de couverture des proportions	80 %	88 %	88 %	95 %
% de couverture des moyennes	88 %	89 %	90 %	94 %

Les résultats de l'étude par simulation sont présentés aux tableaux 3 et 4 et aux figures 5 et 6. Le tableau 3 donne les estimations moyennes des coefficients du modèle et la racine carrée de l'erreur quadratique moyenne (REQM) sur les 300 simulations obtenues en ajustant les trois modèles aux données d'échantillon, soit A- le modèle en deux parties *complet* qui tient compte des corrélations entre les effets aléatoires de district et de village dans les deux parties du modèle, B- le modèle en deux parties ne tenant pas compte des corrélations entre les effets aléatoires de district et de village dans les deux parties, c'est-à-dire en ajustant séparément les deux parties, et C- le modèle linéaire mixte défini par (3), mais ajusté à toutes les réponses, y compris les scores nuls. Ce modèle ne tient pas compte de l'accumulation des scores nuls, mais afin de le rendre plus comparable au modèle en deux parties, nous y avons inclus toutes les variables explicatives comprises dans la partie logistique ou la partie linéaire du modèle en deux parties. Pratiquement, le modèle linéaire mixte ne peut être utilisé que pour prédire les moyennes de district et de village. Pour des raisons de comparabilité, nous avons ajusté les trois modèles en utilisant le logiciel WinBUGS (donc en adoptant le paradigme bayésien), mais il est important de mentionner que l'ajustement des modèles B et C en utilisant le logiciel MLwiN (Rasbash et coll. 2002), qui est beaucoup plus rapide, donne des résultats fort semblables.

Le tableau 3 ne révèle que des différences mineures entre les estimations moyennes et les REQM obtenues en ajustant le modèle complet ou en ajustant les deux parties séparément. Dans le cas de la partie linéaire, les estimations de la moyenne sont très proches des coefficients réels correspondants, ce qui indique qu'il n'y a pas de biais. Pour la partie logistique, les estimations de la moyenne sont de nouveau proches des coefficients réels, quoique les biais estimés soient statistiquement significatifs si l'on se sert de la statistique t classique. Le fait que les REQM sont semblables lorsqu'on ajuste le modèle complet ou les deux parties du modèle séparément donne à penser que, sous les conditions de simulation décrites ici, tenir compte des corrélations entre les effets aléatoires dans les deux parties n'améliore pas l'estimation des coefficients de régression du modèle. Par contre, les résultats du tableau 4 révèlent des biais et des REQM beaucoup plus petits quand on estime les variances du modèle logistique en ajustant le modèle complet, quoique l'estimation de la variance « entre villages » demeure fortement biaisée. L'estimation des corrélations entre les effets aléatoires des deux parties du modèle est satisfaisante. Enfin, comme l'indiquent les deux tableaux, l'ajustement du modèle linéaire mixte, en ne tenant pas compte de l'accumulation des valeurs zéro produit généralement des estimateurs fortement biaisés et, par conséquent, de plus grandes REQM, ce qui n'est évidemment pas étonnant.

La figure 5 montre le biais et la REQM pour la prédiction des moyennes et des proportions réelles de district et de village sous les trois modèles. Soit $\hat{U}_a^r$ n'importe lequel des prédicteurs sous les trois modèles (moyennes ou proportions) pour un domaine donné a tel qu'il est obtenu dans la simulation r et désignons par U_a^r la valeur prédite réelle correspondante. Nous calculons le biais et la REQM comme il suit :

$$\text{Biais}_a = \sum_{r=1}^{300} (\hat{U}_a^r - U_a^r)/300;$$

$$\text{REQM}_a = \left[\sum_{r=1}^{300} (\hat{U}_a^r - U_a^r)^2/300\right]^{1/2}. \quad (12)$$

Les chiffres se rapportant aux villages sont basés sur les 273 villages (parmi les 286) dans lesquels a eu lieu l'échantillonnage. (Comme nous l'avons mentionné plus haut, dans les villages pour lesquels l'ensemble de données original contenait moins de cinq adultes, tous les individus sont inclus dans l'échantillon.)

La conclusion évidente qui se dégage des figures 5a, 5c, 5e et 5g est que l'utilisation du modèle linéaire mixte seulement pour prédire les moyennes de district et de village produit des prédicteurs biaisés dans les domaines échantillonnés ainsi que non échantillonnés, donc de grandes REQM. Notons, cependant, que les REQM des prédicteurs produits sous le modèle linéaire pour les villages sans échantillon sont semblables à celles obtenues sous le modèle en deux parties. Ce résultat est probablement dû au fait que le modèle linéaire mixte est beaucoup plus simple et dépend d'un moins grand nombre de paramètres que le modèle en deux parties, ce qui se traduit, pour les villages pour lesquels il n'existe pas d'échantillon, par des variances de prédiction plus faibles que celles des prédicteurs basés sur le modèle en deux parties. Les figures (5a) à (5d) montrent que les prédicteurs produits sous le modèle en

deux parties, que celles-ci soient ajustées conjointement ou séparément, sont fondamentalement sans biais, malgré le biais dans l'estimation de certains paramètres de la partie logistique observé dans les tableaux 3 et 4. Les figures (5e) à (5h) ne présentent aucune différence appréciable entre les REQM obtenues en utilisant le modèle complet ou en ajustant les deux parties séparément, ce que nous avions constaté également dans les tableaux 3 et 4.

La figure 6 donne le pourcentage de fois que les intervalles de crédibilité à 95 % produits sous les trois modèles couvrent les vraies moyennes et proportions de district ou de village. Voir la section 3.3 pour la construction des bornes des intervalles de crédibilité lorsqu'on utilise des simulations MCMC. La principale conclusion qui se dégage de la figure 6 est que, si l'on ignore l'accumulation de scores nuls et que l'on ajuste le modèle linéaire mixte seulement, on obtient, pour la plupart des domaines, des taux de couverture des vraies moyennes de domaines très différents du taux nominal de 95 %, ces taux étant particulièrement faibles pour les villages pour lesquels il existe un échantillon. L'ajustement du modèle complet produit des taux de couverture un peu meilleurs pour les moyennes de district que l'ajustement des deux parties du modèle séparément, mais les taux de couverture pour les proportions de district sont semblables sous les deux méthodes. Les intervalles de crédibilité pour les moyennes de village semblent différer peu quand on ajuste le modèle complet ou les deux parties séparément, mais il convient de souligner que l'utilisation du modèle complet produit de meilleurs taux de couverture dans 77 % des villages, tandis que l'ajustement des deux parties du modèle séparément ne produit de meilleurs taux de couverture que dans 15 % seulement des villages. Dans les autres villages, les deux méthodes aboutissent aux mêmes taux de couverture. Dans le cas des proportions au niveau du village, les deux méthodes produisent des intervalles de crédibilité semblables, sauf dans quelques cas où le modèle complet donne généralement de meilleurs résultats.

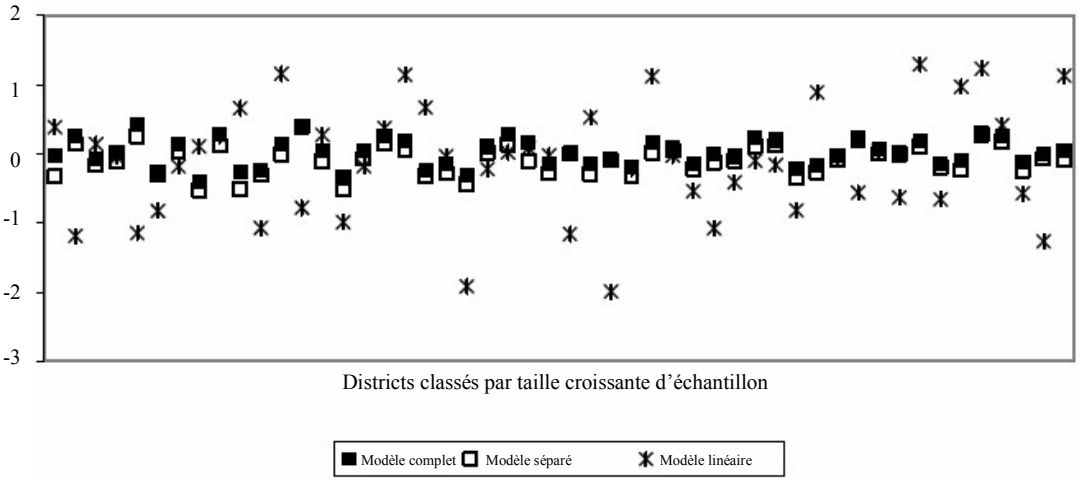
Tableau 3
Moyennes et REQM des estimateurs des coefficients du modèle sous les trois modèles

Coefficient	Vraie valeur	Moyenne par simulation			REQM par simulation		
		Modèle complet	Ajustement séparé	Modèle linéaire	Modèle complet	Ajustement séparé	Modèle linéaire
				Partie linéaire			
β_0	9,38	8,83	9,73	1,90	6,95	6,95	9,21
β_1	4,97	4,97	4,87	12,59	0,32	0,33	7,63
β_2	-1,65	-1,61	-1,58	-3,24	2,05	2,05	2,27
β_3	1,02	1,05	1,05	1,75	0,57	0,57	0,86
β_4	0,94	0,97	0,96	1,51	0,27	0,26	0,60
β_5	-0,01	-0,01	-0,01	-0,02	0,00	0,00	0,01
				Partie logistique			
γ_0	-4,09	-4,38	-4,38	-	0,58	0,59	-
γ_1	1,63	1,73	1,73	-	0,18	0,19	-
γ_2	1,98	2,13	2,13	2,55*	0,41	0,42	7,33*
γ_3	2,06	2,41	2,41	2,05*	0,65	0,65	2,64*
γ_4	-0,35	-0,34	-0,34	0,38*	0,30	0,30	1,37*

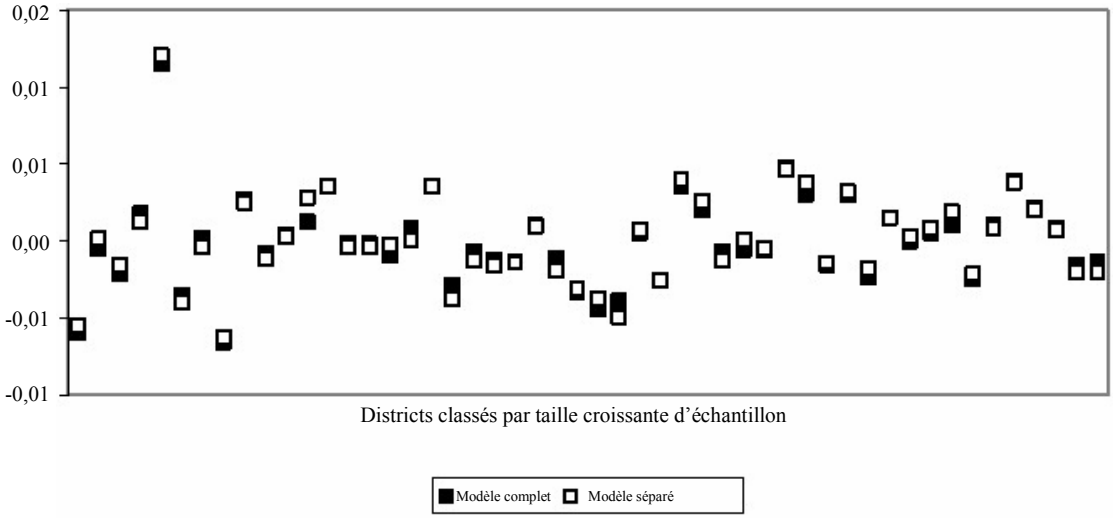
*Estimations obtenues en incluant ces variables explicatives dans le modèle linéaire.

Tableau 4
Moyennes et REQM des estimateurs des variances et des corrélations du modèle sous les trois modèles

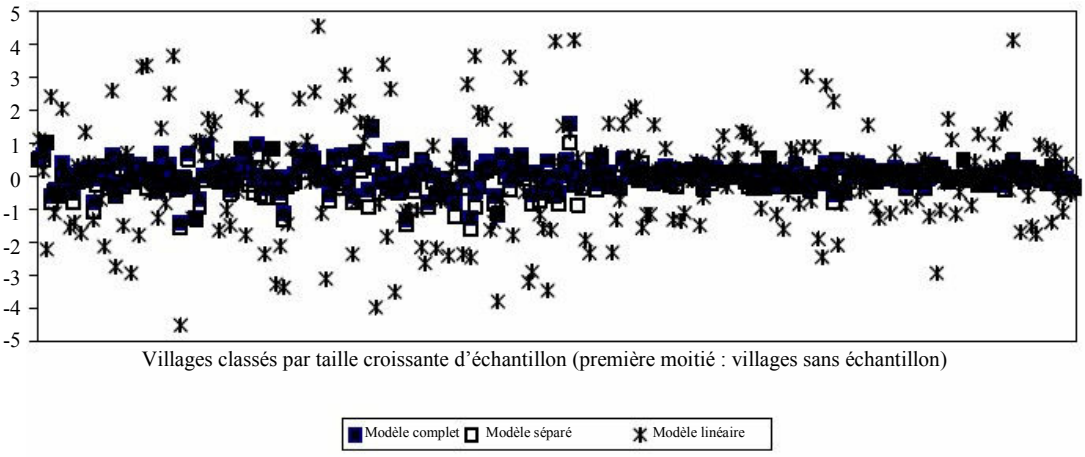
	Vraie valeur	Moyenne par simulation			REQM par simulation		
		Modèle complet	Ajustement séparé	Modèle linéaire	Modèle complet	Ajustement séparé	Modèle linéaire
				Variances – Partie linéaire			
District	60,40	62,23	60,66	103,46	24,52	24,87	46,52
Village	65,44	70,37	70,36	111,97	24,84	25,75	49,74
Résidu	336,00	338,31	338,61	696,82	23,76	24,04	361,64
				Variances – Partie logistique			
District	0,92	1,08	1,50	-	0,61	0,91	-
Village	0,57	0,91	1,15	-	0,70	0,94	-
				Facteurs *K*			
District	0,071	0,075	-	-	0,016	-	-
Village	0,054	0,055	-	-	0,012	-	-
				Corrélations entre les effets aléatoires des deux parties			
District	0,500	0,506	-	-	0,151	-	-
Village	0,500	0,459	-	-	0,148	-	-



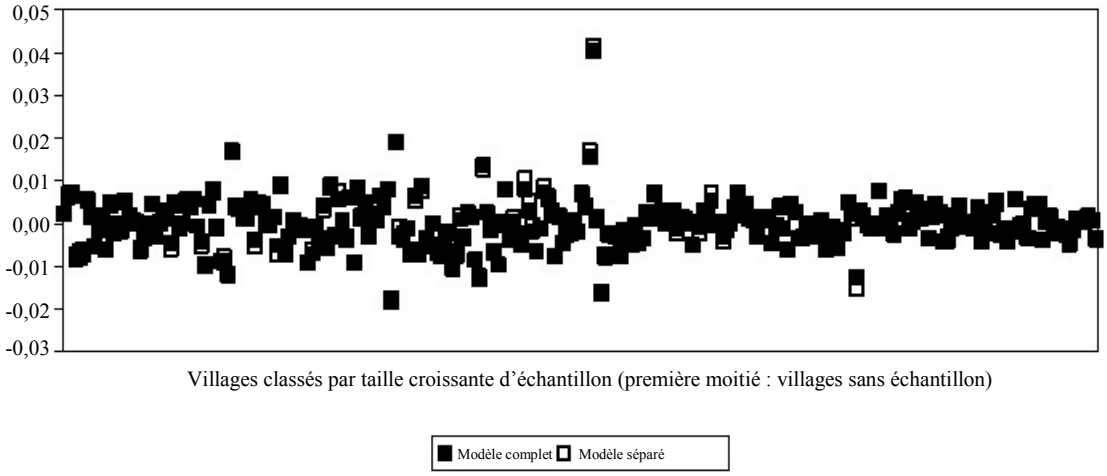
(5a) Biais, moyennes de district



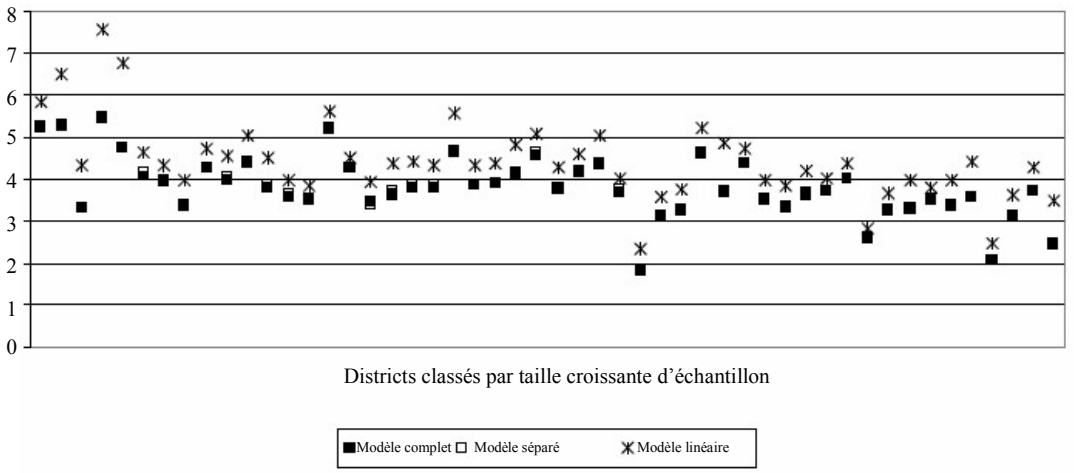
(5b) Biais, proportions par district



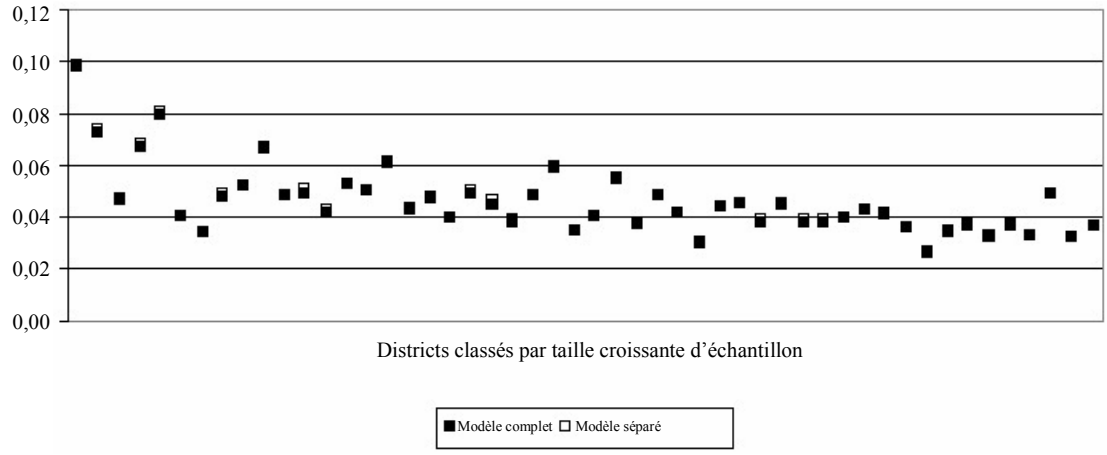
(5c) Biais, moyennes de village



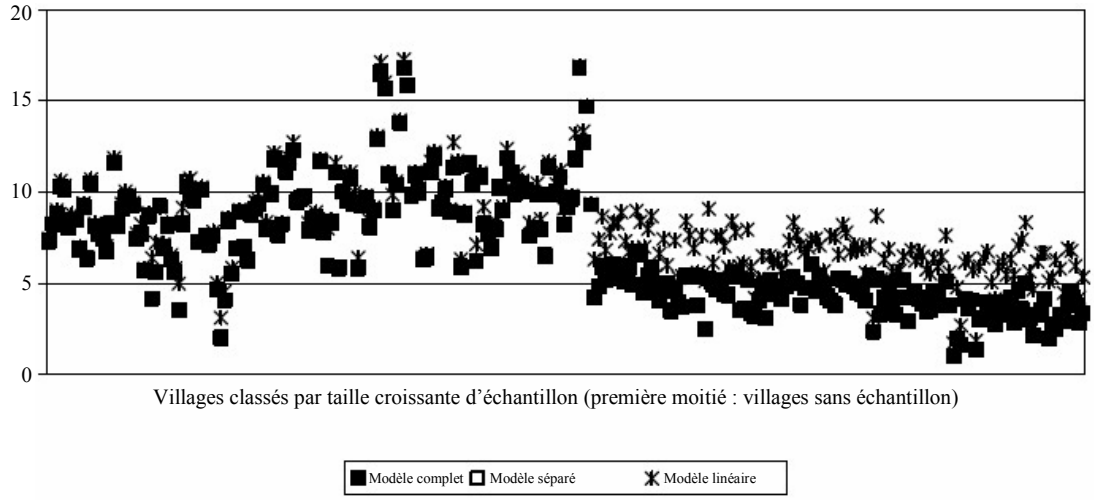
(5d) Biais, proportions par village



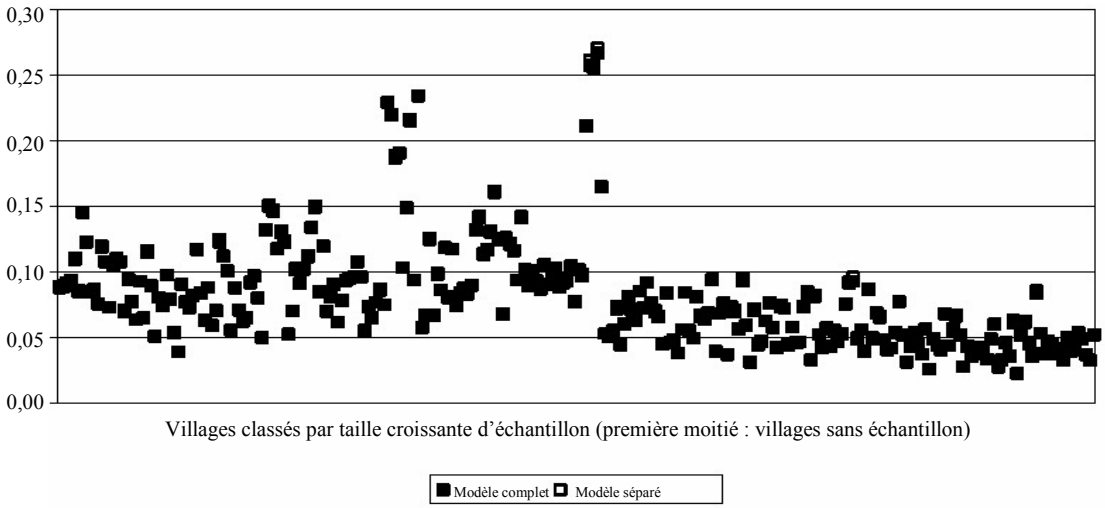
(5e) REQM, moyennes de district



(5f) REQM, proportions par district

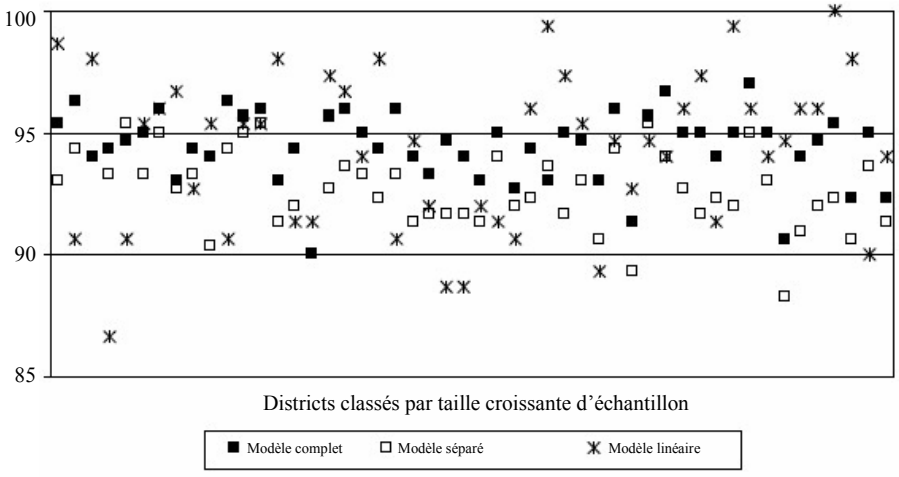


(5g) REQM, moyennes de village

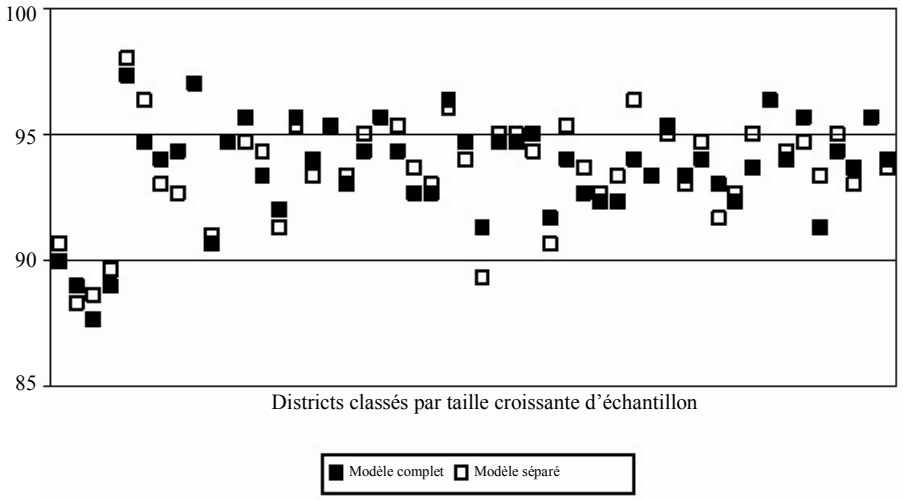


(5h) REQM, proportions par village

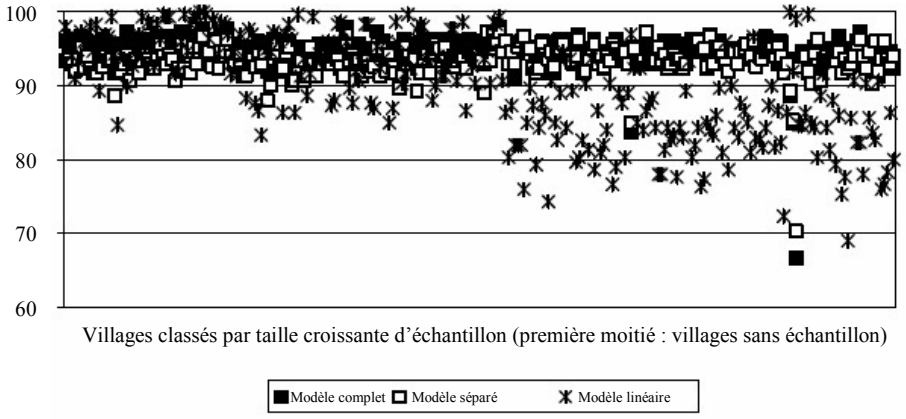
Figure 5 BIAIS et REQM des prédicteurs des moyennes et des proportions par domaine



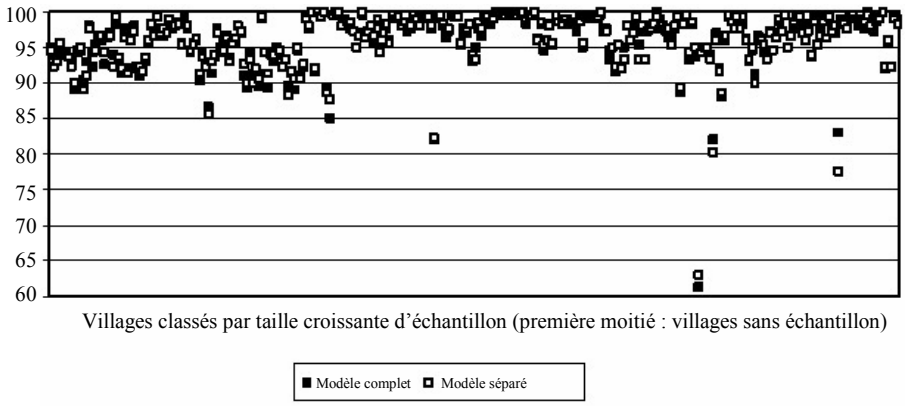
(6a) Moyennes de district



(6b) Proportions par district



(6c) Moyennes de village



(6d) Proportions par village

Figure 6 Taux de couverture des intervalles de crédibilité à 95 % pour les moyennes et les proportions par domaine

4.3 Problèmes de calcul

Comme nous l'avons déjà mentionné, l'ajustement du modèle complet en tenant compte des corrélations entre les effets aléatoires de district et de village dans les deux parties requiert des calculs intensifs et n'est pas toujours stable. En particulier, nous nous sommes heurtés à de sérieux problèmes de calcul lors de l'ajustement du modèle complet avec de très petits échantillons pour la plupart des villages. Par exemple, dans le cas d'un échantillon de 750 individus provenant de 264 villages, tel que, pour près de la moitié des villages, la taille d'échantillon était de 1 ou 2, les valeurs issues des lois a posteriori produites par l'échantillonneur de Gibbs se sont avérés fortement corrélées, même pour de très grands décalages − plus de 1 000 décalages pour les effets aléatoires de village et la corrélation entre les effets aléatoires de village dans les deux parties, et plus de 500 décalages encore après resserrement des lois a priori − si bien que d'extrêmement longues chaînes ont été nécessaires pour obtenir suffisamment de données pour l'inférence. Dans ces conditions, les calculs sont excessivement intensifs et la vérification de la convergence de certaines lois a posteriori est presque impossible. Par conséquent, nous avons sélectionné des échantillons de taille 1 026 dans notre étude par simulation, avec au moins deux individus provenant de chaque village.

5. Résumé

Le message le plus important qui se dégage de la présente étude est que, si l'on ignore l'accumulation de scores nuls et que l'on ajuste un modèle linéaire mixte à l'ensemble de données complet, les prédicteurs peuvent être fortement biaisés et les taux de couverture des intervalles de crédibilité risquent d'être incorrects. Manifestement, la grandeur du biais et la performance des intervalles de crédibilité dépendent, dans ce cas, du pourcentage de scores nuls. L'ajustement d'un modèle en deux parties à ce genre de données produit généralement des prédicteurs sans biais et des intervalles de crédibilité dont les taux de couverture sont acceptables. L'ajustement du modèle en deux parties complet, en tenant compte des corrélations entre les effets aléatoires des deux parties est la meilleure option, mais celle-ci n'améliore que légèrement les prédictions dans notre étude par simulation, malgré l'utilisation de coefficients de corrélation de 0,5 entre les effets aléatoires de district et de village dans les deux parties.

Dans la présente étude, nous avons utilisé des simulations MCMC pour ajuster les modèles et calculer les prédicteurs pour petits domaines et leur variance. Pour appliquer cette approche, il faut spécifier les lois a priori, ce qui peut affecter l'inférence, particulièrement si le nombre de domaines échantillonnés est faible, même en spécifiant des priors non informatifs. Voir Pfeffermann, Moura et Silva (2006) pour une discussion récente et des exemples. L'autre problème lié à l'utilisation de simulations MCMC est la très grande intensivité des calculs. En outre, l'utilisation de cette approche peut devenir instable s'il n'existe que quelques observations dans les domaines échantillonnés. Une autre option consiste donc à ajuster le modèle en deux parties complet suivant l'approche fréquentiste. Les logiciels disponibles comprennent MLwiN (Goldstein 2003) et aML (Lillard et Panis 2003), mais leur utilisation ou celle d'autres logiciels requiert des modifications pour l'estimation de la variance de prédiction en tenant compte des erreurs dans l'estimation des paramètres fixés du modèle. Des méthodes de rééchantillonnage, comme le bootstrap ou le jackknife, pourraient être envisagées pour cela, mais leur adaptation à ce modèle nécessite de nouveaux travaux de développement.

Remerciements

Les travaux de Bénédicte Terryn ont été réalisés en grande partie pendant son séjour à l'Institut de statistique de l'UNESCO (ISU), à Montréal, au Canada. Les travaux de Fernando Moura ont été financés en partie par une subvention de recherche du Conseil national pour le développement de la science et de la technologie du Brésil. Les auteurs remercient un rédacteur associé et deux examinateurs de leurs commentaires constructifs et de leurs suggestions, et Pedro Silva, de son aide pour les calculs finaux.

Bibliographie

Barry, S.C., et Welsh, A.H. (2002). Generalized additive modelling and zero inflated count data. *Ecological Modelling*, 157, 179-188.

Gamerman, D., et Lopes, H.F. (2006). *Markov Chain Monte Carlo*: *Stochastic simulation for Bayesian inference*. 2ième Édition. Chapman & Hall.

Gelfand, A.E., et Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Goldstein, H. (2003). *Multilevel Statistical Models*. 3ième Édition. Londres, Edward Arnold.

Hall, P., et Maiti, T. (2006). On parametric bootstrap methods for small area predictions. *Journal of the Royal Statistical Society*, 68, Séries B, 221-238.

Jiang, J., et Lahiri, P. (2005). Mixed model prediction and small area estimation. *TEST*, 15, 65-72.

Jiang, J., Lahiri, P. et Wan, S. (2002). A unified jackknife method. *Annals of Statistics*, 30, 1782-1810.

Lillard, L.A., et Panis, C.W.A. (2003). *aML Multilevel Multiprocess Statistical Software*, *Version 2.0*. EconWare, Los Angeles, California.

Neal, R.M. (2000). Slice Sampling. Rapport technique Numéro 2005, Department of Statistics, University of Toronto.

Olsen, M.K., et Schafer, J.L. (2001). A two-part random effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96, 730-745.

Pfeffermann, D., Moura, F.A.D.S. et Silva, P.L.D.N. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 943-959.

Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.

Raudenbush, S.W., Yang, M. et Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9, 141-157.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I. et Lewis, T. (2002). A user's guide to MLwiN. Londres : Institute of Education.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Spiegelhalter, D., Thomas, A. et Best, N.G. (2003). *Bayesian Inference using Gibbs Sampling. WinBUGS version 1.4*, *User manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, U.K.

Tanner, M.A. (1996). *Tools for Statistical Inference*, 3ième Édition. New York : Springer-Verlag.

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont fourni de l'aide ou ont fait la critique d'un article ou plus durant l'année 2008.

B. Baffour, *University of Southampton*
J.-F. Beaumont, *Statistique Canada*
Y. Bélanger, *Statistique Canada*
W. Bell, *U.S. Census Bureau*
C. Bocci, *Statistique Canada*
H.J. Boonstra, *Statistique Netherlands*
J. Breidt, *Colorado State University*
P. J. Cantwell, *U.S. Census Bureau*
R. Chambers, *University of Wollongong, Australia*
A. Cyr, *Statistique Canada*
G. Datta, *University of Georgia*
T.J. DeMaio, *U.S. Census Bureau*
J.A. Dever, *University of Maryland*
D. Dolson, *Statistique Canada*
J. Eltinge, *Bureau of Labor Statistics*
M. Ferland, *Statistique Canada*
W.A. Fuller, Iowa State University
J. Gambino, *Statistique Canada*
W. Gamrot, *University of Economics, Katowice, Poland*
D. Garriguet, *Statistique Canada*
M. Ghosh, *University of Florida*
C. Goga, *Institut de Mathématiques de Bourgogne*
S. Haslett, *Massey University, New Zealand*
Y. He, *Harvard Medical School*
D. Haziza, *Université de Montréal/Statistique Canada*
M. Hidiroglou, *Statistique Canada*
N.J. Horton, *Smith College*
J. Jiang, *University of California Davis*
C. Julien, *Statistique Canada*
J. Kim, *Iowa State University*
J.-K. Kim, *Yonsei University, Seoul*
P.S. Kott, *USDA/National Agricultural Statistics Service*
R.A. Kulka, *Abt Associates Inc.*
P. Lahiri, *JPSM, University of Maryland*
H. Lee, *Westat, Inc.*
M. Larsen, *Iowa State University*
P. Lavallée, *Statistique Canada*
J.C. Legg, *Iowa State University*
R. Lehtonen, *University of Helsinki*
S. Lele, *University of Alberta*
C. Léon, *Statistique Canada*
B. Liu, *Westat, Inc.*
S. Lohr, *Arizona State University*
L. Mach, *Statistique Canada*
H. Mantel, *Statistique Canada*
T. Maiti, *Iowa State University*
B. Mandram, *Worcester Polytechnic Institute*
J. Maples, *U.S. Census Bureau*
A. Matei, *Université de Neuchâtel, Suisse*
Y. McNab, *UBC*
T. Merkouris, *Athens University of Economics and Business*
S. M. Miller, *U.S. Bureau of Labor Statistics*
I. Molina-Peralta , *Universidad Carlos III de Madrid*
G.E. Montanari, *University of Perugia*
F. Moura, *Universidade do Brasil-UFRJ*
C. Nadeau, *Statistique Canada*
L. Nordberg, *Statistics Sweden*
J. Oleson, *University of Iowa*
J. Opsomer, *Colorado State University*
M. Park, *Korea University*
M. Pratesi, *Università di Pisa*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université de Laval*
L. Rizzo, *Westat, Inc.*
S. Rubin-Bleuer, *Statistique Canada*
N. Salvati, *Università di Pisa*
F. Scheuren, *National Opinion Research Center*
D. Nóbrega da Silva, *Universidade Federal do Rio Grande do Norte*
S. Singh, *University of Texas at Brownsville*
S. Sinha, *Carleton University*
P. Smith, *Office for National Statistics*
J.-L. Tambay, *Statistique Canada*
Y. Tillé, *Université de Neuchâtel*
M. Torabi, *University of Alberta*
D. Toth, *Bureau of Labor Statistics*
N. Tzavidis, *University of Manchester*
R. Valliant, *JPSM*
F. Verret, *Statistique Canada*
D. Willimack, *United States Census Bureau*
M. Wolfson, *Statistique Canada*
C. Wu, *University of Waterloo*
Y. You, *Statistique Canada*
W. Yung, *Statistique Canada*

Nous remercions également ceux qui ont contribué à la production des numéros de la revue pour 2008: Céline Ethier de la Division de la recherche et de l'innovation en statistique, Christine Cousineau de la Division des méthodes d'enquêtes auprès des ménages, Susie Fortier et Nick Budko de la Division des méthodes d'enquêtes auprès des entreprises, Cécile Bourque, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Denis Coutu, Darquise Pellerin et Isabelle Poliquin (Division de la diffusion), Sheri Buck (Division du développement de systèmes) et Sylvie Dupont (Division des langues officielles et traduction).

PUBLICATIONS ÉLECTRONIQUES DISPONIBLES À
www.statcan.ca

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 24, No. 2, 2008

Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator
Carl-Erik Särndal and Sixten Lundström 167

Model-Based Inference for Two-Stage Cluster Samples Subject to Nonignorable Item Nonresponse
Ying Yuan and Roderick J.A. Little 193

An Implementation Strategy for Efficient Convergence of the Lavallée and Hidiroglou Stratification Algorithm
Patricia Gunning, Jane M. Horgan and Gary Keogh 213

Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure
Natalie Shlomo and Ton De Waal 229

Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation
Mick P. Couper, Eleanor Singer, Frederick G. Conrad and Robert M. Groves 255

Use of Deflators in Business Surveys: An Analysis Based on Italian Micro Data
Leandro D'Aurizio and Raffaele Tartaglia-Polcini 277

An Experiment on Perceived Survey Response Burden Among Businesses
Dan Hedlin, Helen Lindkvist, Helena Bäckström and Johan Erikson 301

Letter to the Editor
Prof. Krishnamurty Muralidhar and Dr. Rathindra Sarathy 319

Book and Software Reviews
Safaa Rabie Amer, John Dixon, Michael D. Larsen, Michael W. Link, Stephen M. Miller, Dr. Jean Ritzen and Mr. Andreas Lindner 323

In Other Journals 339

Contents
Volume 24, No. 3, 2008

Special Section on Longitudinal Surveys 341

Estimating Models for Panel Survey Data under Complex Sampling
Marcel D.T. Vieira and Chris J. Skinner 343

The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study
Ian Plewis, Sosthenes C. Ketende, Heather Joshi and Gareth Hughes 365

Seam Effects in Longitudinal Surveys
Mario Callegaro 387

Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection
Annette Jackle 411

Explaining the Size and Nature of Response in a Survey on Health Status and Economic Standard
Fredrik Johansson and Anders Klevmarken 431

Finite Sample Revision Variances for ARIMA Model-Based Signal Extraction
Tucker McElroy, Richard Gagnon 451

Weight Adjustments for the Grouped Jackknife Variance Estimator
Richard Valliant, J. Michael Brick and Jill A. Dever 469

Book and Software Reviews
Rob Burnside and Charlotte Steeh 489

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 36, No. 1, March/mars 2008

Paul GUSTAFSON
Editor's report 1

Louis-Paul RIVEST
Capture-recapture models: Introduction 3

Saman MUTHUKUMARANA, Carl J. SCHWARZ & Tim B. SWARTZ
Bayesian analysis of mark-recapture data with travel time-dependent survival probabilities 5

Olivier GIMENEZ
Discussion: Towards a Bayesian analysis template? 21

Saman MUTHUKUMARANA, Carl J. SCHWARZ & Tim B. SWARTZ
Authors' response 24

Roger PRADEL, Lory MAURIN-BERNIER, Olivier GIMENEZ, Meritxell GENOVART, Ŕemi CHOQUET & Daniel ORO
Estimation of sex-specific survival with uncertainty in sex assessment 29

Rémi CHOQUET
Automatic generation of multistate capture-recapture models 43

Gilles GAUTHIER & Jean-Dominique LEBRETON
Analysis of band-recovery data in a multistate capture-recapture framework 59

Louis-Paul RIVEST
Why a time effect often has a limited impact on capture-recapture estimates in closed populations 75

Sophie VÉRAN & Jean-Dominique LEBRETON
The potential of integrated modelling in conservation biology:
A case study of the black-footed albatross (Phoebastria nigripes) 85

Kanti V. MARDIA, Gareth HUGHES, Charles C. TAYLOR & Harshinder SINGH
A multivariate von Mises distribution with applications to bioinformatics 99

Qing PAN & Douglas E. SCHAUBEL
Proportional hazards models based on biased samples and estimated selection probabilities 111

Jiahua CHEN, Pengfei LI & Yuejiao FU
Testing homogeneity in a mixture of von Mises distributions with a structural parameter 129

Mayer ALVO
Nonparametric tests of hypotheses for umbrella alternatives 143

Christophe CROUX, Gentiane HAESBROECK & Kristel JOOSSENS
Logistic discrimination using robust estimators: An influence function approach 157

Forthcoming papers/Articles à paraître 175

Acknowledgement of referees' services/Remerciements aux membres des jurys 176

Volume 36 (2008): Subscription rates/Frais d'abonnement 177

Volume 36, No. 2, June/juin 2008

Haiying CHEN, Elizabeth A. STASNY & Douglas A. WOLFE
Ranked set sampling for ordered categorical variables 179

Benjamin KEDEM, Guanhua LU, Rong WEI & Paul D. WILLIAMS
Forecasting mortality rates via density ratio modeling 193

Piotr KOKOSZKA, Inga MASLOVA, Jan SOJKA & Lie ZHU
Testing for lack of dependence in the functional linear model 207

Valeria SAMBUCINI & Ludovico PICCINATO
Likelihood and Bayesian approaches to inference for the stationary point of a quadratic response surface 223

Bruno SANSÓ, Alexandra M. SCHMIDT & Aline A. NOBRE
Bayesian spatio-temporal models based on discrete convolutions 239

Lei SHI & Gemai CHEN
Local influence in multilevel models 259

Sanjoy K. SINHA
Robust methods for generalized linear models with nonignorable missing covariates 277

Zhenlin YANG, Eden Ka-Ho WU & Anthony F. DESMOND
Inference for general parametric functions in Box-Cox-type transformation models 301

Lan WANG
Nonparametric test for checking lack of fit of the quantile regression model under random censoring 321

Forthcoming papers/Articles à paraître 337

Online access to *The Canadian Journal of Statistics*/Services en ligne de *La revue canadienne de statistique* 338

Volume 36 (2008): Subscription rates/Frais d'abonnement 339